frontiers | Research Topics

Perceptual organization in computer and biological vision

Edited by
James Elder, Dirk Bernhardt-Walther, Anitha Pasupathy and Mary A. Peterson

Published in
Frontiers in Computer Science
Frontiers in Psychology

ISSN 1664-8714
ISBN 978-2-8325-5345-9
DOI 10.3389/978-2-8325-5345-9

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Perceptual organization in computer and biological vision

Topic editors

James Elder — York University, Canada
Dirk Bernhardt-Walther — University of Toronto, Canada
Anitha Pasupathy — University of Washington, United States
Mary A. Peterson — University of Arizona, United States

Citation

Elder, J., Bernhardt-Walther, D., Pasupathy, A., Peterson, M. A., eds. (2024). *Perceptual organization in computer and biological vision*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-5345-9

Table of contents

frontiers | Frontiers in Computer Science

TYPE Editorial
PUBLISHED 16 May 2024
DOI 10.3389/fcomp.2024.1419831

OPEN ACCESS

EDITED AND REVIEWED BY
Marcello Pelillo,
Ca' Foscari University of Venice, Italy

*CORRESPONDENCE
Dirk B. Walther
✉ dirk.bernhardt.walther@utoronto.ca

†These authors have contributed equally to this work

RECEIVED 19 April 2024
ACCEPTED 24 April 2024
PUBLISHED 16 May 2024

CITATION
Elder JH, Peterson MA and Walther DB (2024) Editorial: Perceptual organization in computer and biological vision.
Front. Comput. Sci. 6:1419831.
doi: 10.3389/fcomp.2024.1419831

Editorial: Perceptual organization in computer and biological vision

James H. Elder[1,2†], Mary A. Peterson[3,4†] and Dirk B. Walther[5*†]

[1]Department of Electrical Engineering and Computer Science, York University, Toronto, ON, Canada, [2]Department of Psychology, York University, Toronto, ON, Canada, [3]Cognitive Science Program, University of Arizona, Tucson, AZ, United States, [4]Department of Psychology, University of Arizona, Tucson, AZ, United States, [5]Department of Psychology, University of Toronto, Toronto, ON, Canada

KEYWORDS

human vision, computer vision, figure-ground, Gestalt, grouping, perceptual organization

Editorial on the Research Topic
Perceptual organization in computer and biological vision

A principal challenge for both biological and machine vision systems is to integrate and organize the diversity of cues received from the environment into the coherent global representations we experience and use to make good decisions and take effective actions. Early psychological investigations date back more than 100 years to the seminal work of the Gestalt school (Koffka, 1935; Wertheimer, 1938). Yet in the last 50 years, neuroscientific and computational approaches to understanding perceptual organization have become equally important, and a full understanding requires integration of all three approaches (Wagemans et al., 2012; Elder, 2018). Perceptual organization can be defined as the process of establishing meaningful relational structures over raw visual data, where the extracted relations correspond to the physical structure and semantics of the scene. The relational structure may be simple, such as set membership for image segmentation, or more complex, such as sequence representations of contours, hierarchical representations of surfaces, or layered representations of scenes. These structures support 3D scene understanding, object detection, object recognition, and activity recognition, among other tasks.

This Frontiers Research Topic brings together 13 contributions to Frontiers in Psychology and Frontiers in Computer Science, with the aim of presenting a single, unified collection that will encourage integration and cross-fertilization across disciplines. Together, these contributions explore how the brain forms representations of contours, surfaces, and objects over 3D space and time, and the degree to which representations formed by recent deep learning models may be similar or different. Here we briefly introduce these 13 contributions and highlight how they interrelate.

"Shape from dots: a window into abstraction processes in visual perception": What constraints and rules does the visual system use to organize simple visual elements into meaningful contours? Displays of dots provide an interesting way of exploring these grouping rules. Unlike Gabor patches (Field et al., 1993; Kovacs and Julesz, 1993, 1994) and line elements (Pettet, 1999; Drewes et al., 2016; Elder et al., 2018; Baker et al., 2021), dots do not provide orientation information. Nevertheless, observers group dots into contours. Baker and Kellman explore under which geometric conditions people perceive a spatial sequence of dots as executing a smooth vs. abrupt change in orientation. They find that a triplet of dots forming an obtuse angle (more than 90 degrees) is perceived as a smooth contour, whereas a triplet forming an acute angle (<90 degrees) is perceived as an abrupt vertex. Dot displays that describe curvilinear contours as opposed to sharp-angled

vertices allowed for clearer perception, better mental rotation, and more accurate detection of shapes. These results may reflect the underlying statistics of smooth contour curvatures and abrupt orientation changes we encounter in the visual world.

"Combining contour and region for closed boundary extraction of a shape": Ultimately, neural mechanisms must organize local spatial features coded in early stages of the visual system into the coherent object representations we perceive. The grouping cues that support this computation include geometric regularities of the object's bounding contour (e.g., good continuation) as well as photometric regularities within the object (e.g., color similarity; Elder and Zucker, 1996, 1998). In this contribution, Hii and Pizlo propose a foveated shortest-path model of contour grouping to explore the potential fusion of geometric contour and color cues in recovering complete object boundaries. Psychophysical results demonstrate that the human visual system can synergistically combine geometric and color grouping cues, in qualitative agreement with their computational model.

"*Specific Gestalt principles cannot explain (un)crowding*": The contribution from Hii and Pizlo concerned how local elements on the retina are organized into a representation of a coherent figure or object. But perceptual organization extends beyond a single figure to determine how we perceive collections of figures or objects in the scene. One window into this process is provided by the study of *crowding*. Crowding is the phenomenon wherein fine spatial judgements can be made more difficult if extraneous "distractor" elements are brought near to the stimulus being judged. *Uncrowding* refers to the remarkable fact that adding a regular pattern of multiple distractors can release this effect. This phenomenon has generally been attributed to the perceptual organization of these extraneous elements into a perceptual group apart from the stimulus being judged. However, in their psychophysical study, Choung et al. find that, while the degree of uncrowding is strongly correlated with perceived grouping, simple models of perceptual grouping fail to account for this relationship. This suggests that the formation of perceptual groups may depend upon subtle interplays and higher-level perceptual interpretations of the visual stimulus that are not easily captured by a simple combination of Gestalt laws.

"Good continuation in 3D: the neurogeometry of stereo vision": The studies discussed so far provide intriguing insight into perceptual organization in the 2D image plane. But how does this relate to the structure of our 3D visual world? Bolelli et al. note that the back-projected boundaries of solid objects are generally not planar curves (Koenderink, 1984), and their 3D structure can be important to perceptual organization and object understanding. Fortunately, this 3D structure can potentially be recovered via the geometry of the binocular projection. Bolelli et al. introduce a mathematical framework relating the projected geometry of these 3D curves to binocular neural selectivity. Based on tools from sub-Riemannian geometry, their model makes predictions about how interactions between neurons in early visual cortex should depend upon the ocularity and joint position-orientation tuning of the neurons. This model provides a framework for understanding the stereo correspondence problem as well as torsional eye movements.

"The coherent organization of dynamic visual images": The challenge of perceptual organization extends not only over the three dimensions of space but also the dimension of time. The review article by Lappin and Bell details how the brain uses spatiotemporal regularities in moving images to perceptually organize the visual stream into continuous surface representations that support the discrimination of fine spatiotemporal judgements with hyperacuity precision.

"Visual cortical processing—From image to object representation": The foveated shortest-path object grouping model of Hii and Pizlo entails an incremental construction of progressively more global, complex, and complete representations. While Hii and Pizlo do not suggest a specific mapping of their model to brain regions, it is common to assume that such computations proceed hierarchically from early to later visual areas. However, a body of work from Zhou et al. (2000), Craft et al. (2007), von der Heydt (2015), Williford and von der Heydt (2016), and others, provides an alternative account. These findings include neural sensitivity in earlier areas of visual cortex to illusory contours and figure/ground assignment that could only emerge from more global computations, challenging the conventional view. In particular, the identification of *border ownership cells* in cortical area V2 that respond selectively to a contour depending upon the figure/ground sign is strong evidence against a feedforward, hierarchical view of object perception. What is the alternative? von der Heydt reviews computational and neurophysiological research supporting the existence of *grouping cells* (*G* cells) that pre-attentively link neurons in early visual areas that are selective for contours to form representations of global "proto-objects" via recurrent processing. von der Heydt conjectures that these G cells might be located outside of the object pathway in the ventral stream, since recordings in areas V1, V2, and V4 have failed to confirm their existence.

"Backward masking implicates cortico-cortical recurrent processes in convex figure context effects and cortico-thalamic recurrent processes in resolving figure-ground ambiguity": Peterson and Campbell also present evidence against a feedforward account of visual perception. They show that recurrent processing plays an essential role in the perception of classic figure-ground displays that were long taken as evidence that convexity is an important prior in building objects in a bottom-up fashion. Previously, Peterson and Salvagio (2008) and Goldreich and Peterson (2012) found that convexity is a weak figural prior unless it is supplemented by a background prior. The background prior requires homogeneous fill-in concave regions alternating with convex regions. Peterson and Campbell show that the convexity prior and the background prior conflict in traditional displays where both convex and concave regions are homogeneously colored and that recurrent processing resolves this conflict before conscious perception. Furthermore, they identify both cortico-cortical and cortical-thalamic recurrent processes in the perceptual organization of the classic displays. Their experiments show that dynamical recurrent interactions are involved in some of the foundational experiments taken as evidence for a feed-forward model of figure-ground perception.

"Perceptual organization and visual awareness: the case of amodal completion": It has long been debated whether the process of amodal completion of partially occluded objects demands attention and awareness or whether it can occur autonomously. Here, Kimchi et al. report four experiments investigating this

question, using a variant of a color-opponent flicker technique in which a priming stimulus can be presented for a duration necessary for perceptual completion while remaining outside perceptual awareness. Kimchi et al. used this technique to create priming stimuli that cued either a local, global, or ambiguous interpretation of a subsequent target stimulus. They found that when the prime indicated a local completion, local targets were classified faster than global targets, suggesting that local completion can take place without visual awareness. However, when the prime cued a global or ambiguous interpretation, target responses were unaffected by the prime, which they take as evidence that awareness is necessary to resolve ambiguity and to generate a global completion.

"Visual and haptic cues in processing occlusion": Vision is only one of the human senses, and fusion with haptic sensing could be particularly important to inform the perceptual organization of partially occluded objects that are only partially visible to the eye. Prior work has shown that partially occluded faces are more easily recognized when the occluders are stereoscopically rendered to appear in front, rather than behind, the faces. Here, Takeichi et al. use virtual reality to investigate how both visual and haptic information about the relative depth of the occluder affects recognition of katakana characters. While the haptic cue was found to increase the confidence of observer judgements of the relative depth of the occluder, there was no effect on character recognition. Also, counter to prior work with faces, character recognition was better when the "occluder" was rendered to be behind, rather than in front, of the character, suggesting that 3D processing may be different for specialized 2D stimuli like textual characters than for faces.

"The mid-level vision toolbox for computing structural properties of real-world images": The research reviewed above largely follows in the tradition of Gestalt psychology in using highly simplified stimuli to isolate specific perceptual factors and test hypotheses. However, the maturation of computer vision technologies provides opportunity to explore whether principles of perceptual organization generalize to real-world scenes in all of their complexity. Walther et al. provide a useful resource for this endeavor with their Mid-Level Vision (MLV) Toolbox. The toolbox offers algorithms for extracting contours from photographs and for computing a variety of contour properties: orientations, curvature, length, and contour junctions. Relying on the medial axis transform as a dual representation of scene contours, the toolbox provides code to compute measures of local parallelism, local mirror symmetry, and contour separation. The toolbox also contains code for visualizing these properties and for manipulating contour drawings based on them.

"Does training with blurred images bring convolutional neural networks closer to humans with respect to robust object recognition and internal representations?": The success of deep learning models in solving computer vision problems has led to their adoption as potential models for predicting neural and behavioral response to visual stimuli. While these models do capture many aspects of neural and behavioral response, there are intriguing divergences in how networks handle out-of-distribution perturbations such as image blur. Here, Yoshihara et al. find that training convolutional networks with a mix of blurry and sharp images makes them more human-like in their robustness to blur and weighting of shape vs. texture in making classification decisions (Geirhos et al., 2018).

"Shape-selective processing in deep networks: integrating the evidence on perceptual integration": Training with blurred stimuli likely knocks out fine-scale texture cues that networks tend to rely on by default, upweighting the use of shape cues. But what is the nature of the shape cues that these networks can use? While humans make profound use of configural shape information, recent research suggests that deep networks struggle to organize these global shape cues, relying more on local shape features (Baker et al., 2018; Baker and Elder, 2022). In their contribution, Jarvers and Neumann perform a new analysis of deep neural network shape sensitivity that suggests that the addition of recurrent or residual connections can enhance sensitivity to non-local shape, although not to the extent seen in humans. These results suggest future directions for neural network design that may lead to models that are better able to capture the human ability to organize local features into representations of global object shape.

"Self-attention in vision transformers performs perceptual grouping, not attention": Deep learning models have made substantial gains in performance through mechanisms of "self-attention" and "cross-attention" that allow for multiplicative interactions between data inputs and are the basis for more recent state-of-the-art transformer architectures. Here, Mehrani and Tsotsos argue that the effect of self-attention is in fact more appropriately described as *perceptual organization* based on feature similarity. In a series of computational experiments, they demonstrate that vision transformers learn to group stimuli based on features such as hue, lightness, saturation, shape, size, or orientation and suggest that this can be thought of as a form of horizontal relaxation labeling. This novel view provides insight into how transformer architectures may solve difficult perceptual organization problems that challenge convolutional architectures.

Author contributions

JE: Writing—original draft, Writing—review & editing. MP: Writing—original draft, Writing—review & editing. DW: Writing—original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Baker, N., and Elder, J. H. (2022). Deep learning models fail to capture the configural nature of human shape perception. *iScience* 25:104913. doi: 10.1016/j.isci.2022.104913

Baker, N., Garrigan, P., and Kellman, P. J. (2021). Constant curvature segments as building blocks of {2D} shape representation. *J. Exp. Psychol. Gen.* 150:1556. doi: 10.1037/xge0001007

Baker, N., Lu, H., Erlikhman, G., and Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLOS Comput. Biol.* 14:e1006613. doi: 10.1371/journal.pcbi.1006613

Craft, E., Schutze, H., Niebur, E., and von der Heydt, R. (2007). A neural model of figure-ground organization. *J. Neurophysiol.* 97, 4310–4326. doi: 10.1152/jn.00203.2007

Drewes, J., Goren, G., Zhu, W., and Elder, J. H. (2016). Recurrent processing in the formation of shape percepts. *J. Neurosci.* 36, 185–192. doi: 10.1523/JNEUROSCI.2347-15.2016

Elder, J. H. (2018). Shape from contour: computation and representation. *Ann. Rev. Vision Sci.* 4, 423–450. doi: 10.1146/annurev-vision-091517-034110

Elder, J. H., Oleskiw, T., and Fründ, I. (2018). The role of global cues in the perceptual grouping of natural shapes. *J. Vision* 18, 1–21. doi: 10.1167/18.12.14

Elder, J. H., and Zucker, S. W. (1996). "Computing contour closure," in *Proceedings of the 4th European Conference on Computer Vision*, eds. B. F. Buxton and R. Cipolla (Cham: Springer Verlag), 399–412.

Elder, J. H., and Zucker, S. W. (1998). Evidence for boundary-specific grouping. *Vision Res.* 38, 143–152. doi: 10.1016/S0042-6989(97)00138-7

Field, D. J., Hayes, A., and Hess, R. F. (1993). Contour integration by the human visual system: Evidence for a local "association field". *Vision Res.* 33, 173–193. doi: 10.1016/0042-6989(93)90156-Q

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., Brendel, W., et al. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint*. arXiv:1811.12231.

Goldreich, D., and Peterson, M. A. (2012). A Bayesian observer replicates convexity context effects. *Seeing Perc.* 25, 365–395. doi: 10.1163/187847612X634445

Koenderink, J. J. (1984). What does the occluding contour tell us about solid shape? *Perception* 13, 321–330. doi: 10.1068/p130321

Koffka, K. (1935). *Principles of Gestalt Psychology*. Harcourt: Brace and World.

Kovacs, I., and Julesz, B. (1993). A closed curve is much more than an incomplete one: Effect of closure in figure-ground discrimination. *Proc. Natl. Acad. Sci. USA* 90, 7495–7497. doi: 10.1073/pnas.90.16.7495

Kovacs, I., and Julesz, B. (1994). Perceptual sensitivity maps within globally defined visual shapes. *Nature* 370, 644–646. doi: 10.1038/370644a0

Peterson, M. A., and Salvagio, E. (2008). Inhibitory competition in figure-ground perception: context and convexity. *J. Vision* 8, 1–13. doi: 10.1167/8.16.4

Pettet, M. W. (1999). Shape and contour detection. *Vision Res.* 39, 551–557. doi: 10.1016/S0042-6989(98)00130-8

von der Heydt, R. (2015). Figure–ground organization and the emergence of proto-objects in the visual cortex. *Front. Psychol.* 6:1695. doi: 10.3389/fpsyg.2015.01695

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., et al. von der. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychol. Bullet.* 138:1172. doi: 10.1037/a0029333

Wertheimer, M. (1938). "Laws of organization in perceptual forms," in *A Sourcebook of Gestalt Psychology*, ed. W. D. Ellis (London: Routledge and Kegan Paul), 71–88.

Williford, J. R., and von der Heydt, R. (2016). Figure-ground organization in visual cortex for natural scenes. *eNeuro* 3:6. doi: 10.1523/ENEURO.0127-16.2016

Zhou, H., Friedman, H., and von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *J. Neurosci.* 20, 6594–6611. doi: 10.1523/JNEUROSCI.20-17-06594.2000

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 15 February 2023
DOI 10.3389/fpsyg.2023.1047694

OPEN ACCESS

EDITED BY
Dirk Bernhardt-Walther,
University of Toronto, Canada

REVIEWED BY
Shaode Yu,
Communication University of China, China
Ko Sakai,
University of Tsukuba, Japan
Mikio Inagaki,
National Institute of Information and Communications Technology, Japan
Juan Chen,
South China Normal University, China

*CORRESPONDENCE
Shin'ya Nishida
✉ shinyanishida@mac.com

SPECIALTY SECTION
This article was submitted to Perception Science, a section of the journal Frontiers in Psychology

RECEIVED 18 September 2022
ACCEPTED 20 January 2023
PUBLISHED 15 February 2023

CITATION
Yoshihara S, Fukiage T and Nishida S (2023) Does training with blurred images bring convolutional neural networks closer to humans with respect to robust object recognition and internal representations? *Front. Psychol.* 14:1047694. doi: 10.3389/fpsyg.2023.1047694

Does training with blurred images bring convolutional neural networks closer to humans with respect to robust object recognition and internal representations?

Sou Yoshihara[1], Taiki Fukiage[2] and Shin'ya Nishida[1,2]*

[1]Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Kyoto, Japan, [2]NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Atsugi, Japan

It has been suggested that perceiving blurry images in addition to sharp images contributes to the development of robust human visual processing. To computationally investigate the effect of exposure to blurry images, we trained convolutional neural networks (CNNs) on ImageNet object recognition with a variety of combinations of sharp and blurred images. In agreement with recent reports, mixed training on blurred and sharp images (B+S training) brings CNNs closer to humans with respect to robust object recognition against a change in image blur. B+S training also slightly reduces the texture bias of CNNs in recognition of shape-texture cue conflict images, but the effect is not strong enough to achieve human-level shape bias. Other tests also suggest that B+S training cannot produce robust human-like object recognition based on global configuration features. Using representational similarity analysis and zero-shot transfer learning, we also show that B+S-Net does not facilitate blur-robust object recognition through separate specialized sub-networks, one network for sharp images and another for blurry images, but through a single network analyzing image features common across sharp and blurry images. However, blur training alone does not automatically create a mechanism like the human brain in which sub-band information is integrated into a common representation. Our analysis suggests that experience with blurred images may help the human brain recognize objects in blurred images, but that alone does not lead to robust, human-like object recognition.

KEYWORDS

convolutional neural networks, object recognition, visual development, perceptual organization, optical blur

1. Introduction

Human visual acuity, evaluated in terms of the minimum angle of resolution or the highest discernable spatial frequency, is affected by a variety of processes including eye optics, retinal sensor sampling, and the subsequent neural signal processing. In daily visual experiences, visual acuity changes depending on, for example, the degree to which the current focal length of the eye agrees with the distance to the target object, or whether the target is sensed at the fovea, where image sampling is dense, or at far-peripheral vision, where sparse image sampling is followed by spatial pooling. Visual acuity also changes progressively with each stage of development. Infants who are born with low visual acuity gradually acquire near adult-level acuity within the first few

years of life (Dobson and Teller, 1978; Banks and Salapatek, 1981). Considering that the loss of visual acuity can be approximated by blurring the image by low-pass (high-cut) filtering, one can say that most humans have a rich experience seeing blurred visual images in addition to sharp ones.

It has been suggested that the experience of blurred visual images might be functionally beneficial, enabling the visual system to use global configural structures in image recognition (Grand et al., 2001; Le Grand et al., 2004; Vogelsang et al., 2018). Several recent studies test this hypothesis computationally by machine learning using artificial neural networks (Vogelsang et al., 2018; Katzhendler and Weinshall, 2019; Avberšek et al., 2021; Geirhos et al., 2021; Jang and Tong, 2021, 2022). Vogelsang et al. (2018) trained a convolutional network (CNN) to recognize human faces. To simulate how visual acuity gradually improves during the initial stage of life, they changed the training images from blurred to sharp ones during training (B2S) and found that the network achieves robust face recognition for a wide range of image blur, as humans do. In contrast, the network can only recognize sharp images when trained on sharp images. The network can only recognize blurred images when trained using blurred images or sequentially trained on images that change from sharp to blurred. Jang and Tong (2021) found that the effect of B2S training is task-specific. It leads to blur-robust recognition for face recognition as Vogelsang et al. reported, but not for object recognition. Avberšek et al. (2021) was also unable to obtain blur-robust object recognition by means of B2S training. However, object recognition achieves blur robustness when blurred and sharp images are always mixed during training (B+S).

With a similar research motivation in mind, we examined the effects of blur training on object recognition by CNNs. We investigated which types of blur training make the CNNs sensitive to coarse-scale global features as well as fine-scale local features, and bring them closer to the human object recognition system. We evaluated the object-recognition performance of the blur-trained CNNs not only using low-pass filtered test images, but also for other types of images including band-pass filtered images and shape-texture cue conflict images (Geirhos et al., 2019) to ascertain whether blur training affects global configurational processing in general. In agreement with previous reports (Avberšek et al., 2021; Jang and Tong, 2021), our results show that B+S training, but not B2S training, leads to blur-robust object recognition comparable to human performance. However, B+S training is not sufficient to produce robust human-like object recognition based on global configuration features. For example, it reduces the texture bias of CNNs for shape-texture cue conflict images, but the effect is too small to achieve a strong shape bias comparable to that of humans.

In the latter half of this report, using correlation analyses of internal representations and zero-shot transfer learning, we examine how B+S training makes CNN less affected by image blur. Our results suggest that initial low-pass filtering contributes to the blur robustness of B+S-Net, but only partially. Representational similarities in the intermediate layers suggest that B+S-Net processes sharp and blurry images not through separate specialized sub-networks, but through a common blur-robust mechanism. Furthermore, we found that B+S training for other object labels transfers to another label trained only with blurred or sharp images, which suggests that B+S training lets the network learn general blur-robust features. However, blur training alone does not automatically create a mechanism like the human brain where sub-band information is integrated into a common representation.

Overall, our results suggest that experience with blurred images may help the human brain develop neural networks that recognize the surrounding objects regardless of image blurring, but that alone does not lead to robust, human-like object recognition.

2. Methods

We investigated the performance of several training methods with a mixture of blurred images. In the experiments, we mainly used 16-class-ImageNet (Geirhos et al., 2018) as a dataset, and the analysis is based on 16-class-AlexNet, with 16 final layer units. However, we also ran some of the experiments using a 1000-class-ImageNet and tested other network architectures to ensure the generalizability of our results. A list of the networks compared in this study is summarized in Table 1. We trained all the models from scratch except for SIN-trained-Net (Geirhos et al., 2019), for which we used the pre-trained model provided by the authors. We did not fine-tune any of the models for the test tasks. Further, we collected human behavior data via Amazon Mechanical Turk (AMT) to compare human performances with those of our blur-trained models. Below, we provide in detail information about the dataset, model architecture, training strategies, and a human behavior study.

2.1. Dataset: 16-class-ImageNet

In order to facilitate comparison with experimental data on humans, we used the 16-class-ImageNet dataset. This dataset was created by Geirhos et al. (2018), who grouped 1,000 ImageNet classes into superior classes such as “dog" and “clock" and selected the following 16 classes from them: *airplane, bear, bicycle, bird, boat, bottle, car, cat, chair, clock, dog, elephant, keyboard, knife, oven*, and *truck*. There are 40,517 training images and 1,600 test images. There was no overlap between them. The image size is 224 × 224 × 3 (height, width, color). The performance of the model trained on the regular 1000-class-ImageNet is also investigated in a later section (section 3.3.3).

2.2. CNN model: 16-class-AlexNet

We chose AlexNet (Krizhevsky et al., 2012) as the CNN model for our main analysis. We used the model architecture provided in a popular deep learning framework, Pytorch, and trained the model from scratch. To match the number of classes in the 16-class-ImageNet, we changed the output number in the final layer from 1000 to 16.

We chose AlexNet because of its similarity to the hierarchical information processing of the human visual cortex. For example, the visualization of filters in the first layer of AlexNet trained with ImageNet shows the formation of various Gabor-like filters with different orientations and scales (Krizhevsky et al., 2012). The Gabor functions are known to be good approximations of the spatial properties of V1 simple cell receptive fields (Jones and Palmer, 1987). In section 3.3.4, we also analyze a model that explicitly incorporated

TABLE 1 CNN models used in this study.

Model name	Architecture	Number of units in final layer	Training dataset	Pre-training	Fine-tuning
16-class-AlexNet	AlexNet	16	16-class-ImageNet (Geirhos et al., 2018)	No	No
1000-class-AlexNet	AlexNet	1,000	ImageNet(ILSVRC2012)	No	No
VOneNet	VOneBlock (Dapello et al., 2020) + AlexNet	16	16-class-ImageNet (Geirhos et al., 2018)	No	No
16-class-VGG16	VGG16 (Simonyan and Zisserman, 2015)	16	16-class-ImageNet (Geirhos et al., 2018)	No	No
16-class-ResNet50	ResNet50 (He et al., 2016)	16	16-class-ImageNet (Geirhos et al., 2018)	No	No
SIN-trained-Net (Geirhos et al., 2019)	AlexNet	1,000	SIN-ImageNet (Geirhos et al., 2019)	Yes	No

the Gabor filters as the initial layer of AlexNet using VOneBlock proposed by Dapello et al. (2020). A study of the brain hierarchy (BH) score, which takes into account the hierarchical similarity between the deep neural network (DNN) and the brain, shows that AlexNet has a high BH score (Nonaka et al., 2021). The information representation in the convolutional layer of AlexNet corresponds to the lower visual cortex of the brain, while the fully connected layer corresponds to the higher visual cortex of the brain. In addition, AlexNet is an easy model to interpret in that it has a small number of layers and does not contain complex operations such as Skip Connection.

2.3. Training with blurred images: Blur training

In this experiment, in addition to the regular training, we trained CNNs with blurred images using three different strategies (Figure 1). We used Gaussian kernel convolution to blur images. The blur size was manipulated by changing the standard deviation (σ) of the Gaussian kernel as shown in Figure 1A. The spatial extent of the Gaussian kernel (k) was determined depending on σ as follows: $k = \text{Round}(8\sigma + 1)$.[1] When k was an even number, one was added to make it an odd number.

In the following, we refer to the trained models as S-Net, B-Net, B+S-Net, and B2S-Net, respectively, depending on which image blurring strategy was used in training (Figure 1B). Unless otherwise stated, the architecture of each model is 16-class-AlexNet. We trained all the models for 60 epochs (the number of training cycles through the full training dataset), with a batch size of 64. The optimizer was stochastic gradient descent (SGD) with momentum = 0.9 and weight decay = 0.0005. The initial learning rate (lr) was set to 0.01 and decreased by a factor of ten at every 20 epochs. The number of training images was 40,517, the same for all models, and we applied random cropping and random horizontal flipping to all training images. The image size was 224 × 224 × 3 (height, width, color). We used PyTorch (version 1.2.0) and one of two GPU machines to train each model. The GPU environments were Quadro RTX 8000 (CUDA Version: 10.2) and GeForce RTX 2080 (CUDA Version: 10.2).

- **S-Net** is a model trained on sharp (original, unblurred) images.
- **B-Net** In the training of B-Net, all the training images were blurred throughout the entire training period. We mainly discuss the performance of the model trained with a fixed blur size of $\sigma = 4$ px.
- **B+S-Net** In the training of B+S-Net, we blurred half of the samples randomly picked in each batch of training images throughout the entire training period. We mainly discuss the performance of the model trained with a fixed blur size of $\sigma = 4$ px. The performance of B+S-Net trained with randomly varied σ is presented in section 3.3.2.
- **B2S-Net** In the training of B2S-Net, the training images were progressively made sharper from a strongly blurred to the original, non-blurred image. Specifically, we started with a Gaussian kernel of $\sigma = 4$ px and decreased σ by one every ten epochs so that only sharp images without any blur were fed into the model in the last 20 epochs. This training method is intended to simulate human visual development and to confirm the effectiveness of starting training with blurred images, as claimed by Vogelsang et al. (2018).

To ensure the reproducibility of the results, we trained each network models with eight different initial weights, and computed the mean and the 95% confidence intervals for each condition.

2.4. Human image classification task

We collected human data using Amazon Mechanical Turk (AMT). We asked participants to perform an image classification task to investigate the difference between the models trained in this study and human image recognition capabilities.

As stimuli for the classification task, we used the same 16-class-ImageNet test set that we used for evaluating CNN models (1,600 images, 100 images per class). In addition to the original test images, we tested the low-pass and band-pass versions of the 16-class-ImageNet test images for stimuli. The low-pass images were created by applying Gaussian kernel convolution while manipulating

1 The equation is implemented in OpenCV's GaussianBlur function that we used to apply Gaussian filtering. In this function, the kernel size was adaptively determined from the size of sigma.

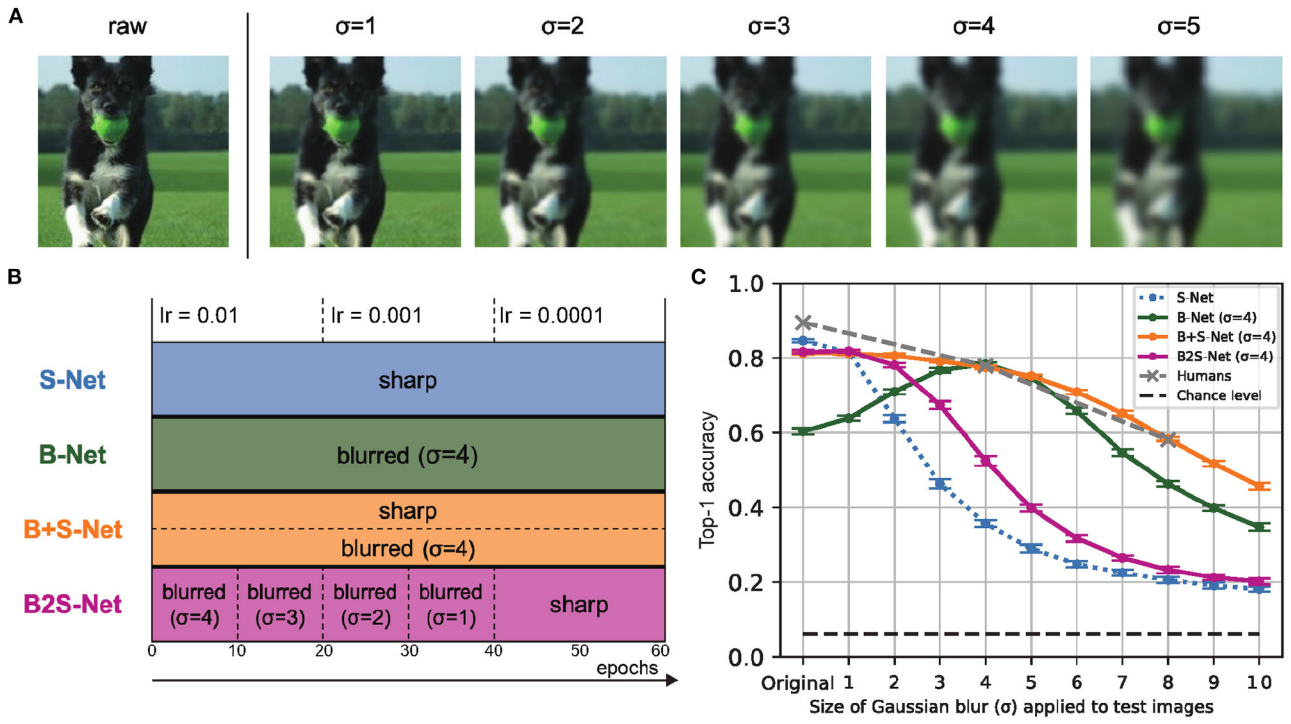


FIGURE 1
Blur training brings CNN closer to humans in recognizing low-pass filtered object images. **(A)** Sample blurred images. **(B)** Four blur training methods. **(C)** Classification accuracy on low-pass filtered test images. Top-1 accuracy is the rate at which the model's first choice matches the expected answer. Error bars represent 95% confidence intervals estimated from the performances of eight models trained with different random seeds.

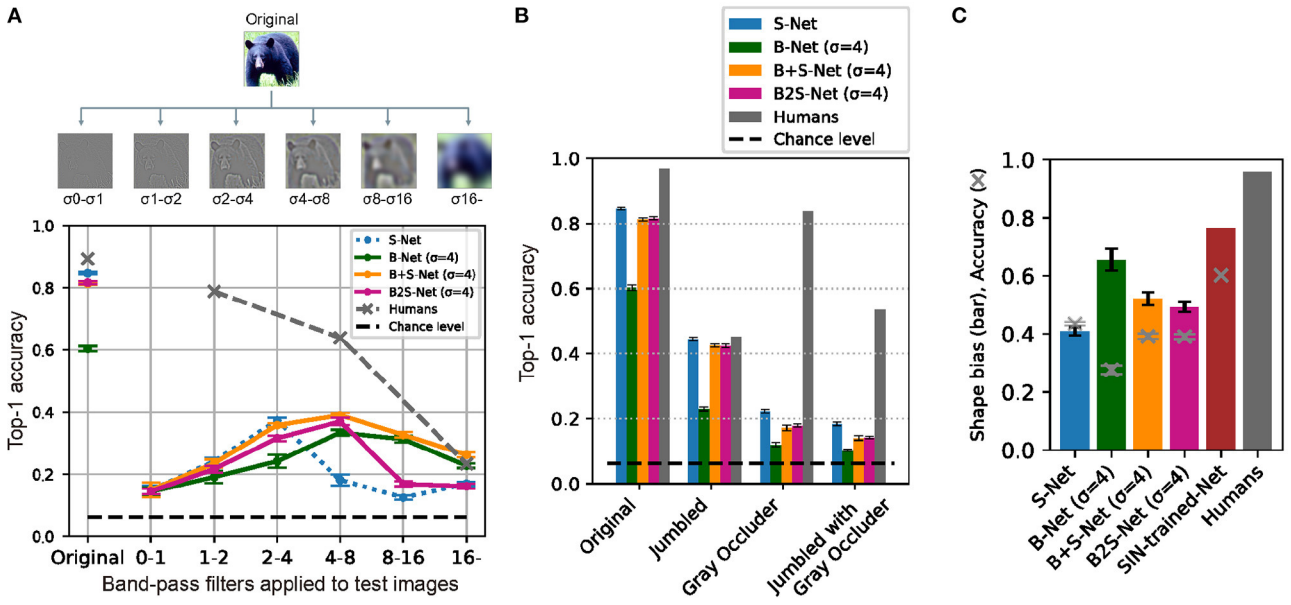


FIGURE 2
Blur training does not bridge the gaps between CNN and humans in recognition of **(A)** band-pass filtered object images, **(B)** jumbled/occluded images (Keshvari et al., 2021), and **(C)** shape-texture-cue-conflict images (Geirhos et al., 2019). The ordinate is the Top-1 accuracy of object classification. The exception is **(C)**, where shape bias is shown by bars, and classification accuracy on Stylized ImageNet (SIN) by crosses. Human accuracy for **(B, C)** are from the original studies. Error bars represent 95% confidence intervals estimated from the performances of eight models trained with different random seeds.

the standard deviation of the Gaussian kernel (σ) in the same manner as when we blurred the training images for CNN models. The band-pass images were created by taking the difference between two low-pass images obtained by blurring the same image with different σ (Figure 2A). In total, there were six conditions as follows: original image, low-pass image $\sigma = 4$ px, low-pass image $\sigma = 8$ px, low-pass image $\sigma = 16$ px, band-pass image $\sigma 1 - \sigma 2$ (i.e., band-pass image obtained by subtracting the lowpass image with

$\sigma = 2$ px from that with $\sigma = 1$ px), and band-pass image $\sigma 4 - \sigma 8$.

For each task, one of the stimuli was presented and participants chose the category of an object in the image from 16 options. We had 170 subjects solve the tasks and obtained 6,817 pieces of categorization data (Original image: 1,108, low-pass image $\sigma = 4$: 1,103, low-pass image $\sigma = 8$: 1,124, low-pass image $\sigma = 16$: 1,134, band-pass image $\sigma 1 - \sigma 2$: 1,188, band-pass image $\sigma 4 - \sigma 8$: 1,160). Participants could complete the task for an arbitrary number of images. The consent form for the experiment was created using a Google Form, and a link to it was placed on the AMT task page. Each participant was asked to read the linked consent form and fill in the required information to register his or her consent. Experimental procedures were approved by the Research Ethics Committee at Graduate School of Informatics, Kyoto University, and were conducted in accordance with the Declaration of Helsinki.

3. Blur training: Results

We measure the model's performance on various test images and compare it to human performance to investigate what visual functions are acquired via blur training. First, we examined the classification accuracy for low spatial frequency images and analyzed whether the models could recognize coarse-scale information. Then, we examined whether the robustness to blurry images acquired through blur training could generalize to other types of robustness measured by using band-pass filtered images, images with manipulated spatial configurations of local elements, and shape-texture cue conflict images.

3.1. Recognition performance for low-pass images

In this section, we compare the image recognition performance for low spatial frequency images. Since the low-frequency information can capture global image features to some extent, the results of this task are expected to indicate, at least partially, whether the model recognizes global information or not. For this purpose, we examined the percentage of correct classifications for each model when the test image was blurred at different intensities. The test images are the test set of the 16-class-ImageNet containing 1,600 images. We also collected human classification task data using the same test images. The details are described in section 2.4.

The results of the above experiments are shown in Figure 1C. First, S-Net trained only on standard clear images shows a sharp drop in the accuracy when the image is strongly blurred. The B-Net's performance is high only for the blur level used in training ($\sigma = 4$) and blurs of similar strength. B2S-Net did not show much improvement in blur tolerance.

On the other hand, in the low-frequency image recognition test, B+S-Net, which was trained on both blurred and sharp images simultaneously, was able to recognize a wide range of features from sharp images to blurred images of various intensities (blur robustness). The robustness of B+S-Net against blur is similar to that indicated by human behavior data.

B2S-Net showed only a tiny improvement in terms of accuracy over S-Net, and B+S-Net showed stronger blur robustness than B2S-Net.

3.2. Recognition performance for other types of image manipulations

In the previous section, we showed that the models trained on both low-pass-filtered and sharp images acquire robustness to a broad range of image blur strengths. To gain a more detailed insight into what visual functions were acquired by blur training, we investigated the behavior of blur-trained models for image manipulations that were not used in the training procedure.

3.2.1. Recognition performance for band-pass images

First, we used band-pass images to investigate the recognition performance of the model in each frequency band. The band-pass images were created by subtracting the two low-pass images of different σ. Using these band-pass images, we were able to find out in which frequency band blur training is influential. We also analyzed whether there is a difference between CNNs and humans regarding the frequency bands they can use for object recognition.

In our experiments, we performed a classification task on band-pass images for CNN models and humans, respectively. In the human experiments, we used AMT to conduct the image classification task for band-pass images $\sigma 1 - \sigma 2$ and $\sigma 4 - \sigma 8$. The details are described in section 2.4.

The results (Figure 2A) show that B+S-Net improves the accuracy of image recognition over a broader frequency range than the other models, and this indicates that training on blurred images is effective in acquiring the ability to recognize a broader range of frequency features. However, it did not show much effect on images in the high-frequency band.

Next, we compared the accuracy of humans and CNNs. The CNN model showed a lower recognition rate for band-pass stimuli, especially in the high-frequency range. Blur training does not lead to robust, human-like object recognition for bandpass images.

3.2.2. Recognition performance for global configuration made of local patches

We further investigated whether blur training could change the global information processing in CNN models by using the test procedure proposed by Keshvari et al. (2021).

Keshvari et al. (2021) tested the difference in recognition performance between humans and CNNs by manipulating local patches. They divided the original image into several square tiles. There were four partitioning scales: (4×4), (8×8), (16×16), (32×32). A *Jumbled* image was one in which tiles were randomly replaced horizontally, preserving local information but distorting global shape and configural relationships. The *Gray Occluder* image, in which tiles were alternately grayed out, preserved the global shape and configural information but lost some local information. The *Jumbled with Gray Occluder* image combined the operations of the *Jumbled* and *Gray Occluder* images, and both local and global information were destroyed.

Keshvari et al. (2021) compared the difference in recognition accuracy between a CNN (VGG16 pre-trained on ImageNet without blur traning) and human observers using 640 images from eight classes in ImageNet. They found the pretrained CNN showed a significant decrease in accuracy for the *Jumbled* image, and the larger decreases for the *Gray Occluder* image, and the *Jumbled with Gray Occluder* image. Humans also showed a similar magnitude of decrease in accuracy for the *Jumbled* image, but only a small decrease for the *Gray Occluder* image. Their findings suggest that humnans can, but the pretrained CNN cannot, make use of global configural information preserved in the *Gray Occluder* image for object recognition.

In this study, we generated the jumbled/occluded images from the test images of the 16-class ImageNet in the same way as in Keshvari et al. (2021) and investigated whether the recognition performance of CNNs becomes closer to that of humans by blur training (Figure 2B). The results showed that training on blurred images did not change the overall trend of recognition performance on this test set. B+S-Net and B2S-Net did not improve the accuracy for *Gray Occluder* images compared to S-Net. These results suggest that the CNN models failed to utilize the global configural information preserved in the *Gray Occluder* image even after blur training.

3.2.3. Recognition performance for texture-shape cue conflict images

To investigate whether the blur-trained models show a preference for shape information or texture information, we tested the shape bias proposed by the work of Geirhos et al. (2019).

Geirhos et al. (2019) created a texture-shape cue conflict image dataset where the texture information of one image was replaced by that of another image in a different class by using the style transfer technique of Gatys et al. (2016).[2] The dataset consists of the same 16 classes as in the 16-class-ImageNet while each image has two correct labels based on its match to the shape or texture class. In total, the dataset contains 1,200 images (75 images per class). The shape bias measures how often the model answers the shape class when it correctly classifies a cue conflict image into either the shape or texture class, and is calculated by the following equation:

$$shape\ bias = \frac{correct\ shape\ decisions}{correct\ shape\ decisions + correct\ texture\ decisions.}$$

According to the results of Geirhos et al. (2019), while humans showed strong shape bias, CNN models trained on ImageNet showed weak shape bias (in other words, they showed texture bias). When the CNN models were trained on the Stylized-ImageNet (SIN) dataset, in which the texture information of an image was made irrelevant to the correct label by replacing the original texture with that of a randomly selected painting, the shape bias of the CNN models (SIN-trained-Net) became closer to that of humans. Moreover, we found SIN-trained-Net has a higher recognition rate for high-pass and band-pass images as humans do (Supplementary Figure S2B). However, training with SIN is biologically implausible and therefore not helpful in modeling the development of the human visual system.

Here, we calculated the shape bias of the models trained in our study using the texture-shape cue conflict image dataset provided by the authors of Geirhos et al. (2019) to see whether the blur training could enhance the shape bias of CNNs. Figure 2C presents the shape bias of the four models we trained as well as those of SIN-trained-Net and human data taken from Geirhos et al. (2019). Compared to S-Net, shape bias was increased most for B-Net, the second for B+S-Net, and the least for B2S-Net. However, the classification accuracy on the SIN dataset was significantly decreased for B-Net, only slightly for B+S-Net, and not at all for B2S-Net. Overall, among the four models, B+S-Net shows the most human-like performance. However, neither B-Net nor B+S-Net shows strong shape bias comparable to those of SIN-trained-Net and humans. These results indicate that while training with blurred images slightly increases the shape bias in comparison with training only with sharp images, blur training alone is insufficient to bring the bias closer to the human level.

3.3. Supplemental analyses

3.3.1. Training schedule

We used the fixed schedule of learning rates as shown in Figure 1B. We determined the learning rate following a reference training script in torchvision library: https://github.com/pytorch/vision/tree/main/references/classification. To check the generality of our findings in particular about B2S-Net, we have additionally run a supplemental experiment to examine the effect of the training schedule. We trained B2S models while varying the initial learning rate and the number of epochs with discrete step sizes of [0.05, 0.01, 0.005, 0.001] and [60, 90, 120], respectively. The initial learning rate was reduced by a factor of 10 for every third of the total training epochs (as in the original experiment). The timing to decrease the sigma of the Gaussian kernel applied to the training images was also linearly extended (decreasing the sigma by 1 every 10, 15, and 20 epochs for the 60, 90, and 120 epoch training conditions, respectively).

As a result, we have obtained qualitatively similar amounts of blur robustness for all tested conditions except for a model with the initial learning rate = 0.05 and with training epochs = 120, in which the training diverged due to too large initial learning rate. All the models trained with learning rates 0.01 and 0.005 showed blur robustness/accuracy equivalent to the original B2S model regardless of the training epochs. We did not find any model that significantly outperformed the blur robustness of the original B2S model.

3.3.2. B+S-Net with randomly varying blur strength

Considering that B2S-Net simulates human visual experiences during development, one can also consider that B+S-Net simulates human visual experience in everyday life where blurred images are occasionally mixed with sharp images due to image focusing errors. In the analysis so far, we have fixed the strength of image blur applied to training images for B+S-Net at $\sigma = 4$. Here, we trained a 16-class B+S-Net while randomly varying σ (0 px–4 px) to simulate our daily visual experience more realistically, and measured its performance on the (A)low-pass images, (B) jumbled/occluded images, and (C) shape-texture-cue-conflict images.

2 Texture-shape cue conflict image: taken from the GitHub page of Geirhos et al. (2019): https://github.com/rgeirhos/texture-vs-shape/tree/master/stimuli/style-transfer-preprocessed-512, reference date: 2021/07/26.

The results (Supplementary Figure S1) showed no significant changes in the performance on any of the test sets from the original B+S-Net. Fixing the blur strength is not the reason why blur learning is limited in its ability to reproduce human-like global object recognition.

3.3.3. 1000-class-AlexNet

The analysis so far has been based on the 16-class-AlexNet. One may consider that 16 object classes are unrealistically small to simulate human object recognition. To address this concern, we also trained our networks with a 1000-class classification task (1000-class-AlexNet). For comparison with the main results, we used the 16-class-ImageNet to test performances,[3] by mapping the output of the 1000-class-AlexNet into 16 classes based on WordNet hierarchy (Miller, 1995) using the mapping function described in Geirhos et al. (2018).

Concerning the accuracy for blurred images (Supplementary Figure S2A), the 1000-class-AlexNet exhibited an overall trend similar to that of the 16-class-AlexNet. However, we also found that the generalization effect of blur training beyond the blur strength used in training was smaller for the 1000-class-AlexNet than that for the 16-class-AlexNet. B-Net was firmly tuned to the blur strength used in training ($\sigma = 4$) and was barely able to recognize clear images. B+S-Net also showed a narrower blur tuning. B2S-Net showed no advantage over S-Net. Concerning the performances on the band-pass-filtered test images (Supplementary Figure S2B), the results of the 1000-class-AlexNet showed a similar trend to the 16-class-AlexNet. The effective bandwidth was somewhat narrower in the 1000-class version. It should be also noted that the 1000-class-AlexNet trained on Stylyzed-ImageNet (Geirhos et al., 2018) showed a human-like performance for band-pass test images. The results of the shape bias using the cue conflict images (Supplementary Figure S2D) show that there was little effect of blur training on shape bias when the 1000-class dataset was used. However, it should also be noted that the accuracy of the 1000-class models for the cue conflict images themselves was very low, meaning that the models were barely able to classify the test images to either the correct shape or texture label in the first place.

To conclude, we found no evidence supporting the idea that increasing the number of training categories makes blur training more effective in reproducing human-like robust object recognition.

3.3.4. VOneNet (16-class)

VOneNet is a model in which the first layer of the 16-class-AlexNet is replaced with a VOneBlock (Dapello et al., 2020). The VOneBlock is a computational model that simulates the visual information processing in the V1 cortex of the brain, such as the response properties of simple cells and complex cells. It also simulates the stochasticity in neural responses by introducing noise. Importantly, multiscale Gabor filters tuned to low to high spatial frequencies are hard-coded in the VOneBlock.

One possible reason for the limited effect of blur training in reproducing human-like robust object recognition is that the training cannot produce human-like multi-scale filters in the early processing stage. If this were the case, through blur training, the model with VOneBlock would be able to achieve stronger robustness to low-pass and band-pass filtered images and stronger sensitivity to global configurations.

Contrary to this expectation, the introduction of the VOneBlock did not change the performance significantly. As shown in Figure 3, the results for each test set showed a remarkable degree of similarity between the models with and without VOneBlock. Thus, changing the lower-level layer to a model closer to the visual cortex did not affect the effects of blur training in terms of frequency and shape recognition.

3.3.5. VGG16, ResNet50

Finally, we examine the performance of different network architectures other than AlexNet. The networks studied here are VGG16 (Simonyan and Zisserman, 2015) and ResNet50 (He et al., 2016). In general, the results were similar to those obtained with AlexNet, while the performance tended to be more tuned to trained blur strength (Supplementary Figures S3, S4).

4. Analysis of the internal representation of B+S Net

Thus far, we have analyzed the effect of training with blurred images on the basis of recognition performance, and found that the recognition performance of B+S-Net for low spatial frequency images is similar to that of humans. We have focused on the behavioral similarity/dissimilarity between humans and neural nets, leaving the internal processing of the B+S Net as a black box. In this section, we attempt to understand how B+S-Net acquires blur robustness similar to humans by analyzing internal representation analysis. The question is whether B+S-Net processes sharp and blurry images in a way computationally similar to the human visual system.

In general, when a visual processing mechanism is able to recognize both sharp and blurry images, we believe the way the image signals are processed inside the system can be roughly categorized into two cases.

- **Case 1:** Sharp and blurry images are processed by a common general process. Representations for sharp and blurry images are integrated into a common feature representation in the early to middle stage of visual processing. The following information processing is shared (Figure 4, top).
- **Case 2:** The sharp and blurred image features are processed separately by stimulus-specific processes until the outputs of the separate processes are integrated at the last stage to recognize the object (Figure 4, bottom).

Although we know no direct empirical evidence, it is likely that the structure of the human visual system for blurry image recognition is closer to Case 1 than to Case 2. This is because computational resource is more efficiently used in Case 1 than in Case 2. Considering that the human visual system has to cope with a wide range of image deformation other than image blur, having an efficient processing structure with a common higher stage must be a reasonable choice. On the other hand, CNNs with powerful learning abilities may create a specialized sub-network, each processing blurred images and sharp

3 The models trained on the 1000-class dataset were directly used in the analysis. Fine-tuning the 16-class ImageNet may yield different results.

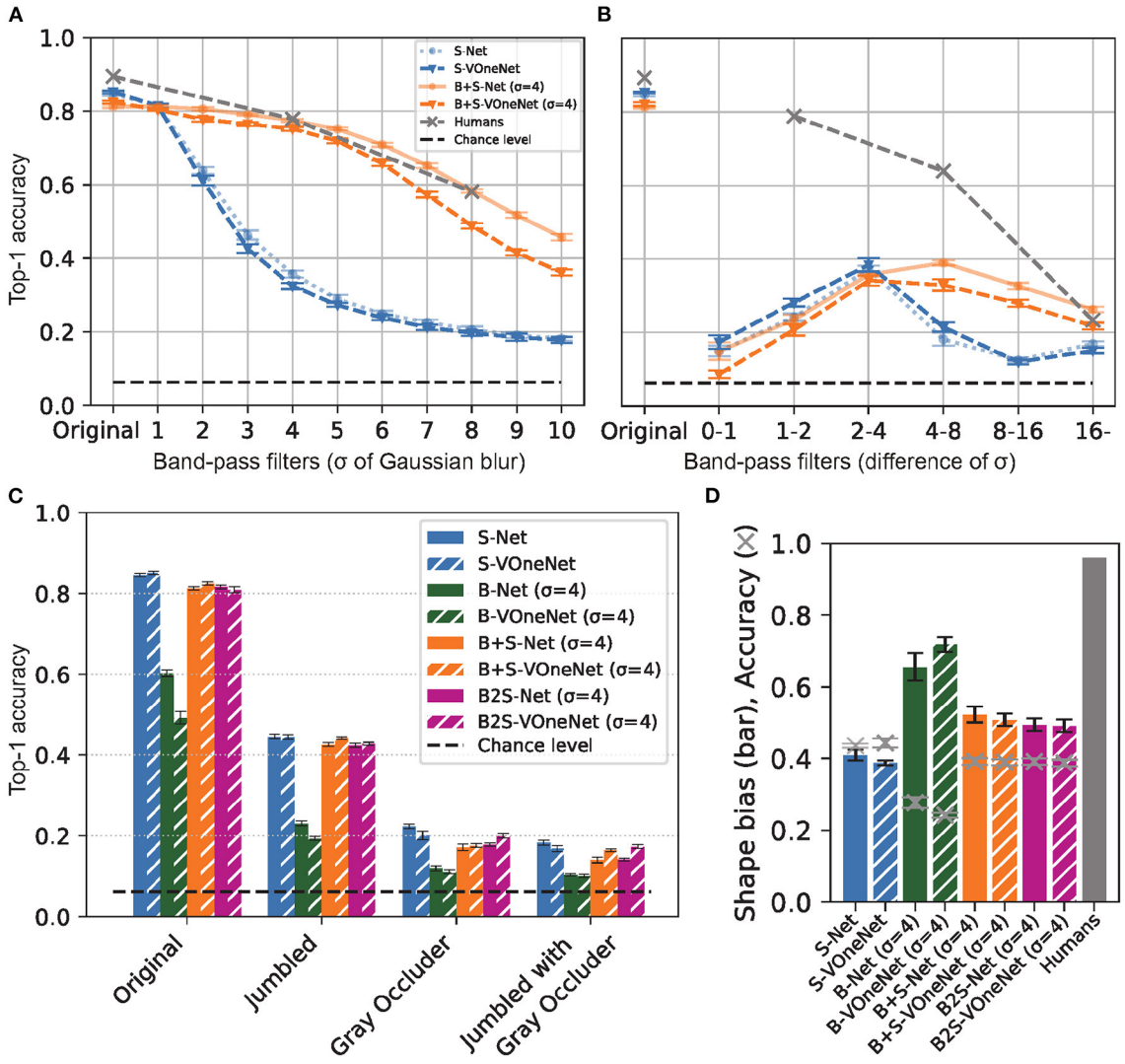


FIGURE 3
VOneNet with blur training. **(A)** Low-pass filtered object image recognition. **(B)** band-pass filtered object image recognition, **(C)** jumbled/occluded image recognition, and **(D)** shape-texture-cue-conflict image recognition. Although VOneNet has fixed V1-like first-stage mechanisms, the effects of blur training are similar to those of AlexNet. Error bars represent 95% confidence intervals estimated from the performances of eight models trained with different random seeds.

images separately, to optimize performance. B+S-Net could be a hybrid of B-Net and S-Net with little interactions between them. When using CNNs as a computational tool to understand human-like robust processing, we should check whether the processing strategy CNNs use to achieve blur robustness is not dissimilar to that of humans. If it were found to be dissimilar, we could learn little about the internal processing of the human visual system from this research strategy.

4.1. Receptive fields in the first layer

First, we visualized the receptive field of the first convolutional layer (Supplementary Figure S5) as was done in previous studies (Vogelsang et al., 2018; Jang and Tong, 2021). Some receptive fields look similar to those found in the early visual cortex. It appears that training method slightly alters the receptive fields. B+S-Net shows a shift of the spatial frequency tuning to the lower frequency compared to S-Net. In other words, B+S-Net is more sensitive to low-frequency information in the first layer.

Low-pass filtering is one way to make the internal representations similar between sharp and blurry images. How much the change in the spatial frequency tuning affects the representational similarity for sharp and blurry images in the first layer will be quantitatively evaluated in the next section.

When 1000-class-AlexNet is compared with 16-class-AlexNet, features with higher spatial frequencies are extracted. This may be because 1000-class-AlexNet needed to extract finer local features to perform more fine-grained classification. This tuning difference may explain why 1000-class-AlexNet shows weaker blur robustness than 16-class-AlexNet.

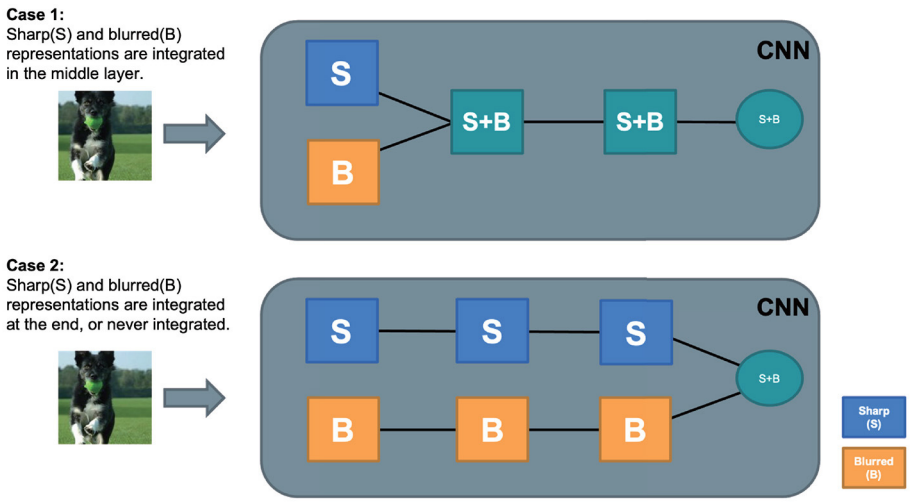


FIGURE 4
Two hypotheses about intermediate feature representation of B+S net. Case 1 is more efficient and presumably close to human processing.

4.2. Correlation of activity in the intermediate layers

To analyze how the sharp and blurred image features are processed in each layer of the CNN models, we computed the average correlations of unit activities in each intermediate layer between the sharp and blurred image inputs (S-B correlation). The more representations and processing shared between clear and blurred images, the higher will be the activity correlation within the layer. Here, we calculated the S-B correlations in the following three cases: (1) the sharp and blurry image pair is generated from the same image, (2) different images from the same class, and (3) different images from different classes. For each case, we computed correlations for all possible sharp-blurry image pairs from 1,600 test images of the 16-class ImageNet. Then, the correlations were averaged across image pairs. The unit activities after the ReLU activation function were used to compute the correlations. (1) is for evaluating the representational similarity at the image level, while (2) and (3) are for evaluating the representational similarity at the category level. By comparing these three, we can infer both representational similarities and the corresponding processing stages.

Figure 5 presents the S-B correlation in each layer of 16-class S-AlexNet (left) and B+S-AlexNet (right).

In the initial layer (Conv1), while S-B correlations are close to zero when different images of the same class (the broken orange line) or different classes (the dotted green line) are used, they are high when the sharp-blurry image pairs from identical images are used (the solid blue line). When S-Net and B+S-Net are compared, B-S correlations are slightly higher for B+S-Net (0.84) than for S-Net (0.76). This agrees with the change in spatial frequency characteristics of the receptive filed we observed in the last section. If low-pass filtering in the first layer were powerful enough to completely remove the difference between sharp and blurry images, the correlation would be one.

In the subsequent convolutional layers, S-B correlations remain to be close to zero when different images of the same class or a different class are used. In S-Net, S-B correlation for the same images gradually drops as the layer goes. This suggests that these layers reduce the representational similarity between sharp and blurry images by extracting fine-scale image features only available in sharp images. On the other hand, in B+S-Net, S-B correlation for the same image remains high. This suggests that these layers extract robust image features commonly available in sharp and blurry images, supporting the idea that B+S-Net achieves blur-robust recognition by forming a common internal processing structure consistent with Case 1.

In the final full-connection layers, S-B correlations gradually increase for the same image and for the same class, while increasing and then dropping for the different class. The pattern of change is similar for S-Net and B+S-Net, but the correlations for the same image/class are higher for B+S-Net, in agreement with the higher classification accuracy of B+S-Net for both sharp and blurry images.

To see the generality of our finding, we also applied the same analysis to 1000-class AlexNet (Supplementary Figure S6) and 16-class VOneNet (Figure 6). In general, the patterns of B-S correlations for both are similar to that we found for 16-class AlexNet, but two issues are worth mentioning. First, S-B correlation in the first convolutional layer is lower for 1000-class AlexNet than for 16-class AlexNet (0.65 for S-Net and 0.69 for B+S-Net), in agreement with the higher-frequency preference of the initial receptive fields for 1000-class AlexNet (Supplementary Figure S5). Second, for VOneNet in which the first layer is hard-coded as a Gabor filter bank, while S-B correlation in the first convolutional layer is the same for S-Net and B+S-Net, S-B correlation of B+S-Net elevates in the subsequent layers. This indicates that B+S-Net forms the features common to both sharp and blurred images from the multiband information extracted in the first layer. The initial low-pass filtering is effective, but not necessary for B+S-Net to achieve blur-robust object recognition.

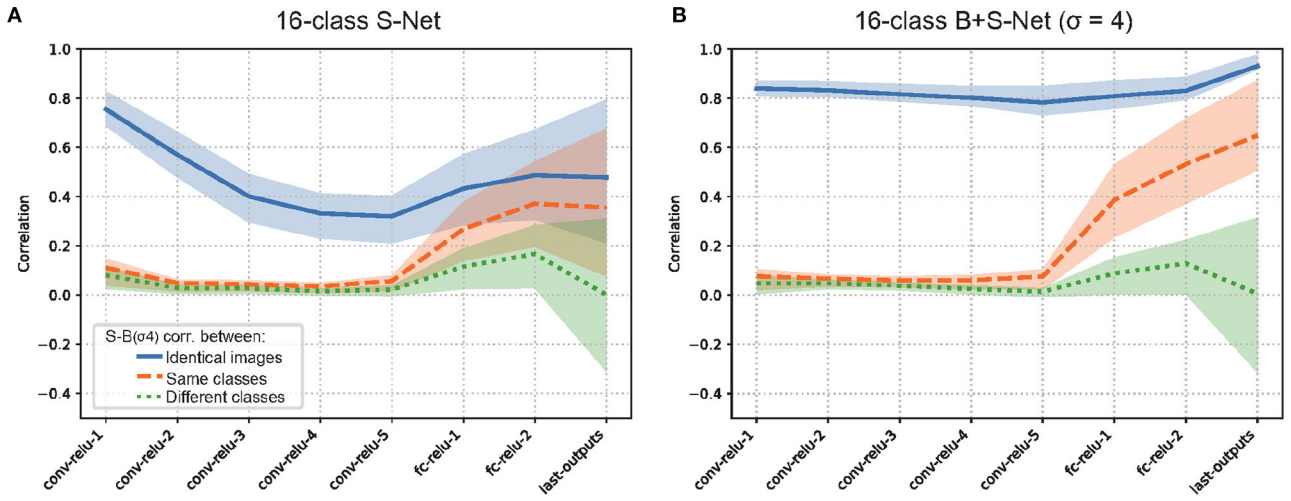


FIGURE 5
Representational similarity of sharp (unblurred) and blurred image inputs (S-B correlation) for S-Net **(A)** and B+S Net **(B)**. Blue: Pearson correlation in unit activity between sharp and blurred versions of the identical images. Orange: Correlation between sharp and blurred versions of different images of the same object class. Green: Correlation between sharp and blurred versions of different images of different object classes. The average correlation of the units in the layer with the interquartile range (25%-75%) is shown. 16-Class AlexNet. The results are consistent with Case 1.

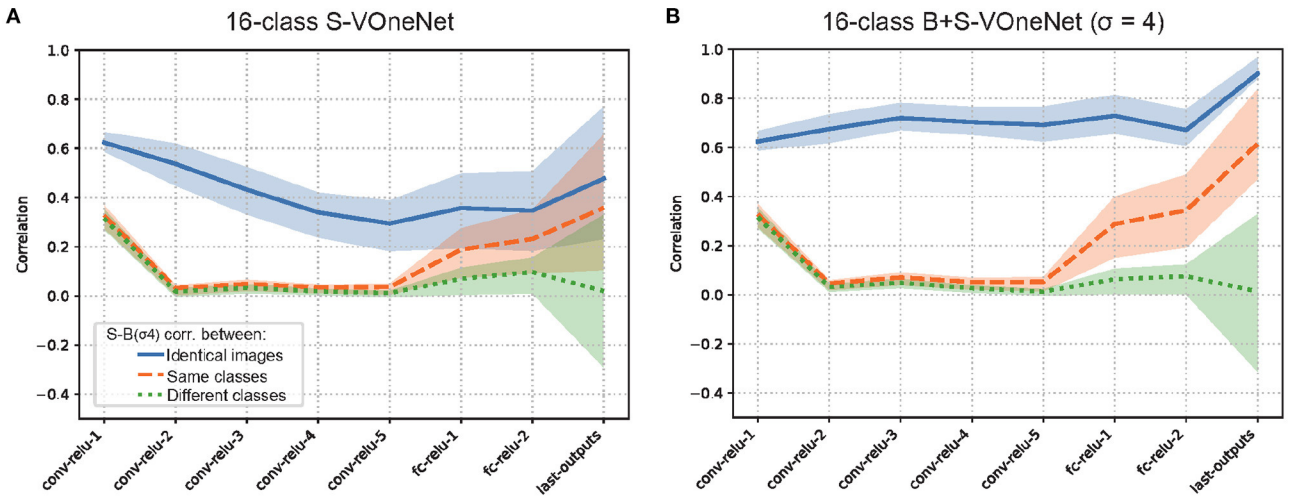


FIGURE 6
Representational similarity of sharp (unblurred) and blurred image inputs for S-Net **(A)** and B+S Net **(B)**. VOneNet. The pattern of results is similar to Figure 5.

4.3. Visualization of the internal representations by t-SNE

To understand how the sharp and blurry images are represented in the intermediate layers of the CNN models, we also attempted to visualize them using the dimensionality reduction algorithm, t-SNE (van der Maaten and Hinton, 2008). Specifically, we recorded the activities of each layer obtained from sharp and blurry images and compressed them into two dimensions for visualization. The two input parameters of t-SNE, perplexity and iteration, were set to 30 and 1000, respectively. The results shown here are visualizations of 10 pairs of sharp and blurred images of the same image sampled for each of the 16 classes.

The visualizations of the intermediate layer activities for the sharp and blurred images are shown in Supplementary Figures S7, S8. First, in early convolutional layers of S-Net, the representations of the sharp and blurry images overlap, and those of the same class are scattered. As the layer goes deeper, the representations of the sharp and blurry are separated, and only sharp images of the same class are clustered. Blurry images remain scattered and separated from sharp images in the final output. Next, in early convolutional layers of B+S-Net, the representations of the sharp and blurry images overlap, and those of the same class are scattered, as in S-Net. As the layer goes deeper, however, representations of the sharp and blurry images do not separate, and both sharp and blurry images of the same class are clustered. These results agree with the

trends indicated by the representational similarity analysis in the last section, providing further support of the idea that B+S-Net achieves blur-robust recognition by forming a common internal processing structure consistent with Case 1.

4.4. Generalization test using zero-shot transfer learning

The results of the S-B correlation analysis in sections 4.2 and 4.3 suggested that the representations are shared between the sharp and blurry versions of the same images in the intermediate layers of B+S-Net. This suggests that the intermediate layers of B+S-Net may have the ability to extract robust image features effective in recognizing both sharp and blurry images. One way to test this idea is to see whether the B+S training extends its effect beyond the image classes used in training, since general robust features should be useful in general.

Using zero-shot learning, we examined the generalizability of the shared representation acquired by blur training to the unseen classes. We trained a subset of object classes, either one or eight in 16 classes, without using blurry images while training the remaining classes using both blurry ($\sigma = 4$) and sharp images, and later evaluated the classification accuracy for that subset of classes using blurry images. Conversely, we also trained a subset of classes without using sharp images while training the other classes using both blurry and sharp images, and later evaluated the classification accuracy using sharp images. Therefore, there were in total four conditions, i.e., training without blurry or sharp images for one or eight classes (w/o 1/16B, w/o 8/16B, w/o 1/16S, w/o 8/16S). We used the 16-class AlexNet for this test.

The recognition accuracy for the unseen image types (either blurry or sharp) in the four test conditions is shown in Table 2.

When either blurry or sharp images were excluded for half of the training classes, the models were not able to recognize these classes of images with the unseen image type. On the other hand, when either blurry or sharp images were excluded for one training class, the accuracy for the unseen class is about three times the chance level ($\frac{1}{16} = 0.0625$). Therefore, although the effect of the generalization of the sharp and blurry features to unseen categories was limited in terms of the zero-shot transfer performance, some amount of transfer was clearly observed at least when there was only one excluded class.

To further analyze the internal representations of the models trained in the transfer experiment, we examined the S-B correlations in the intermediate layers of each model. When either blurry or sharp images were excluded for half of the training classes (Figures 7B, D), the S-B correlation from identical images is significantly reduced in the middle to high layers for the unseen category (orange line), compared to that for the seen category (blue line). On the other hand, when either blurry or sharp images were excluded for one of the training classes (Figures 7A, C), the S-B correlation from identical images for the unseen category (orange line) remains almost as high, albeit slightly lower than that for the seen category (blue line). Therefore, although the shared representation for the blurry and sharp images did not seem to generalize well to the unseen class in terms of the performance level, the similarity of the internal representations appeared to be high between the seen and unseen classes for the model with one excluded class. The reason for this apparent discrepancy is presumably because the misclassification to a class with a similar representation was induced by the imperfect alignment of blur-sharp representations. In fact, a confusion matrix (Supplementary Figure S9) indicates that the misclassifications in the model with one excluded class were mostly from "No.15: truck" class to "No:6 car" class.

Overall, the zero-shot transfer analysis suggested that the shared representation acquired by blur training can be reused, at least partially, to recognize an object class with an unseen image type (either blurry or sharp) during training. This further supported the view that common representations that are invariant to blurry and sharp image inputs are formed in the early and middle stages of visual processing by blur training (Case 1 in Figure 4). In a similar way, humans might efficiently acquire blur robust representations to general object categories just by being exposed to blurry images of a limited number of objects.

4.5. Internal representations for high-pass and low-pass images

We have analyzed the internal representation of B+S-Net for sharp and blurry images, and found B+S-Net has an efficient human-like processing mechanism at least for these images. However, we have also shown in section 3 that B+S-Net does not behave similarly to humans in object recognition for other modified images including high-pass filtered images. To further evaluate B+S-Net as a computational model of the human visual system, we analyzed its internal representation for high-pass and low-pass (blurry) images. A recent human fMRI study (Vaziri-Pashkam et al., 2019) suggests that the representations for high-pass and low-pass images of the same object category are segregated in V1, while integrated and clustered in the higher visual areas.

To investigate how the high- and low-frequency information is represented in S-Net and B+S-Net, we visualized the activity in the intermediate layers for 10 pairs of high-pass (H, $\sigma 1 - \sigma 2$) and low-pass (L, $\sigma = 4$) versions of the same image using the t-SNE (van der Maaten and Hinton, 2008) (perplexity = 30, iteration = 1,000). The visualization results (Supplementary Figures S10, S11) show that the representations of high-pass and low-pass images are less segregated in B+S-Net than in S-Net. We cannot find class-based clustering of high-pass and low-pass images in higher layers of either S-Net or B+S-Net, in agreement with our finding in section 3 that neither S-Net nor B+S-Net can recognize objects in high-pass images, but in disagreement with the representation in human visual cortex (Vaziri-Pashkam et al., 2019).

To further examine the representations for high-pass and low-pass images, we computed the average activity correlation (H-L correlation) of the middle layers of S-Net and B+S-Net between the high- and low-frequency images (Figure 8). In the convolutional layers, the H-L correlation was low even for the same image. Slightly higher correlations for B+S-Net than for S-Net suggest that early layers of B+S-Net have more broadband tuning. In the fully connected layers, the H-L correlation gradually increased. Although this is in line with class-based clustering of high-pass and low-pass images, the increasing trend was weak, and was not enhanced by blur training. The average same-class correlation did not exceed 0.3 for the final output of B+S-Net. In sum, there are significant differences

10.3389/fpsyg.2023.1047694

TABLE 2 The results of zero-short learning test.

Training method	Unseen labels	Seen labels
Training without B images for one class (w/o 1/16B)	0.19	0.78
Training without B images for half of the classes (w/o 8/16B)	0.05	0.87
Training without S images for one class (w/o 1/16S)	0.15	0.81
Training without S images for half of the classes (w/o 8/16S)	0.03	0.90

Classification accuracy for B+S-Net trained without using blurred or sharp images for specific object classes. Chance level accuracy = 0.0625(= $\frac{1}{16}$).

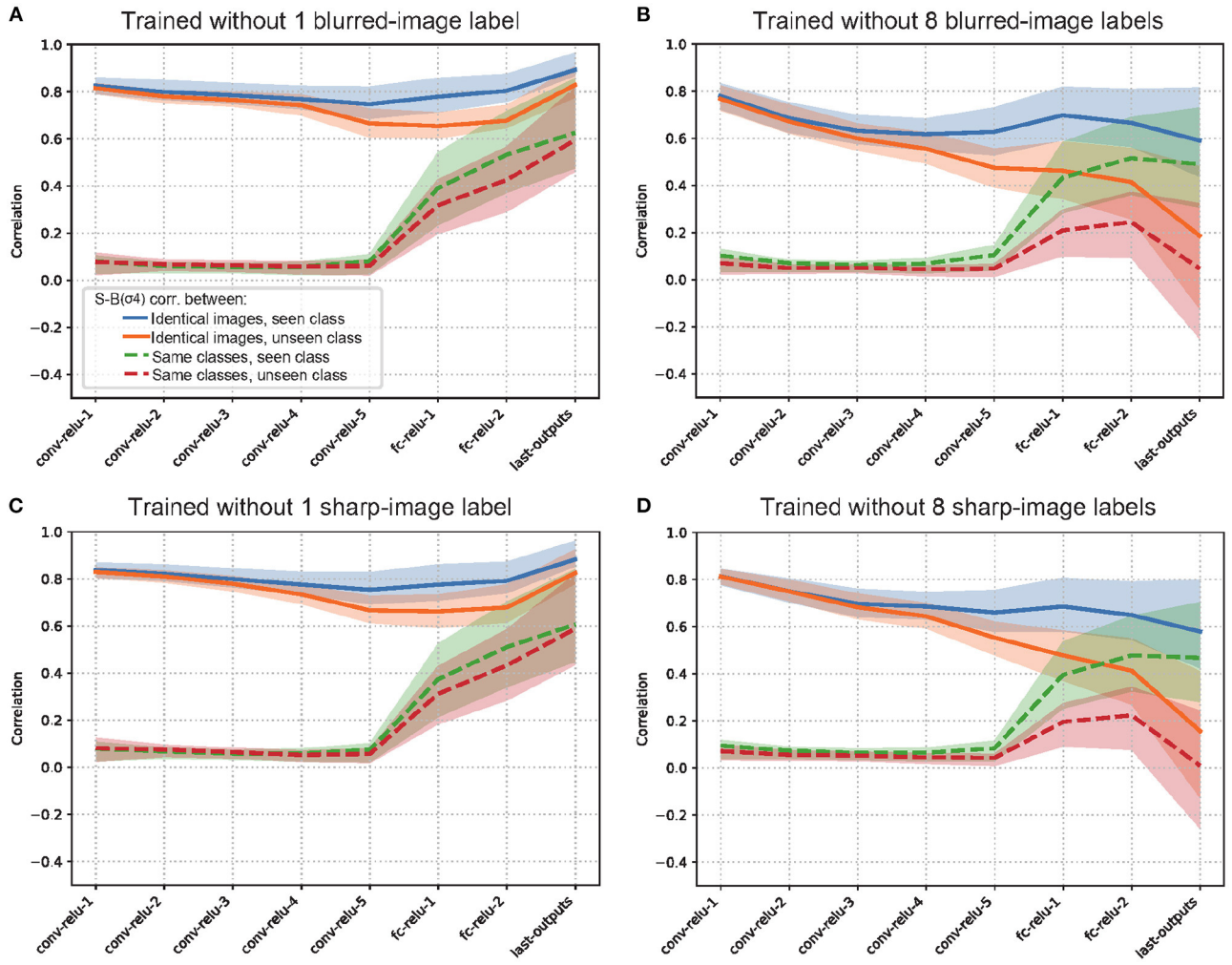


FIGURE 7
Representational similarity analysis for zero-shot learning test when **(A)** one blurred-image label, **(B)** eight blurred-image labels, **(C)** one sharp-image label, or **(D)** eight sharp-image labels is/are excluded from the training set. Test set was blurred images for **(A, B)**, and sharp images for **(C, D)**. The results suggest the shared representation acquired by blur training can be reused, at least partially, to recognize an unseen object class.

in the internal representations for high-pass and low-pass images between B+S-Net and the human cortex, and there is no evidence that blur training facilitates high-level frequency integration.

To see the effect of initial layer on the representation of high-pass and low-pass images, we also analyzed the H-L correlation of VOneNet, which has fixed multi-scale Gabor filters in the first layer. The H-L correlation for the same images in the convolutional layers is low, and, again, there is no evidence of strong integration of low- and high-frequency information in higher layers, unlike representation in the human visual cortex (Vaziri-Pashkam et al., 2019).

5. General discussion

In this study, we investigated the effect of experiencing blurred images on forming a robust visual system to the environment as one of the factors for constructing an image-computable model of the human visual system. To this end, we compared the recognition performance of CNN models trained with a mixture of blurred images using several different strategies (blur training). The results show that B+S-Net trained with a mixture of sharp and blurred images is the most tolerant of a range of blur and the most

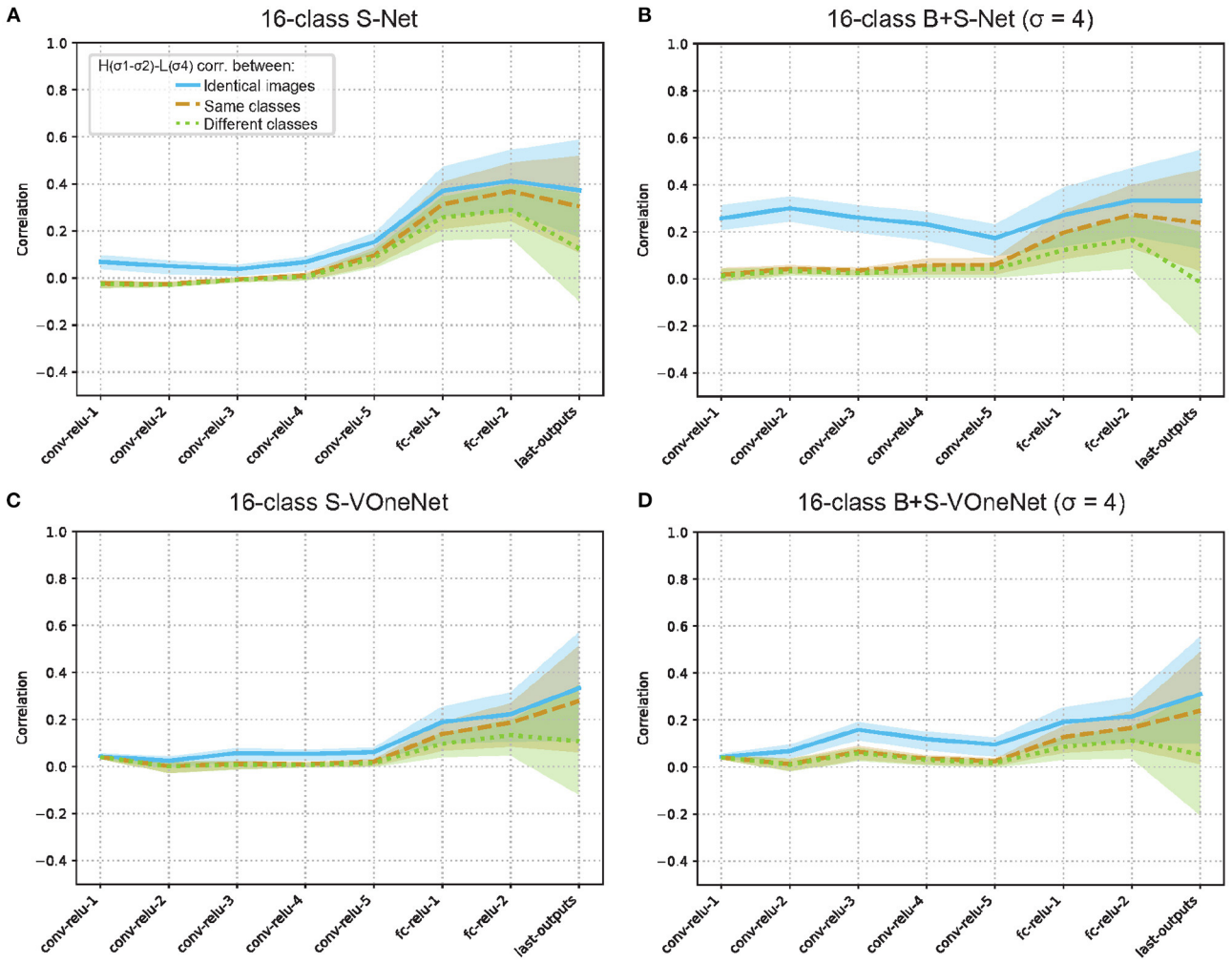


FIGURE 8
Representational similarity of low-pass and high-pass image inputs for S-Net **(A, C)** and B+S Net **(B, D)**. 16-Class AlexNet **(A, B)** and VOneNet **(C, D)**. There is no evidence that blur training facilitates high-level frequency integration as found in human visual cortex.

human-like. In addition, training in the order from blurred to sharp images was not very beneficial. Other evaluations of the model's performance with test stimuli showed that blur training did not improve the recognition of global spatial shape information, or only slightly. The analysis of the internal representation suggests that B+S-Net extracts common features between sharp and blurred images. However, it does not show integration of multi-scale (high and low frequency) frequency information, unlike in the human visual cortex.

In section 3, we compared the effect of training with blurred images on the CNN models in terms of object recognition performance. In all the CNN models we tested, the recognition performance of low spatial frequency features was improved by blur training. In particular, the model trained simultaneously on blurred and sharp images (i.e., B+S-Net) showed blur robustness across a wide range of image blur close to that of humans.

On the other hand, the blur robustness of B2S-Net was weaker than that of B+S-Net. The models showed better performance when trained on blurred and sharp images simultaneously, rather than on a schedule that simulated human visual development. This result apparently disagrees with the study of Vogelsang et al. (2018), which showed that training in the order of low resolution to high resolution improved blur robustness of a CNN model in face recognition. This difference can be attributed to the difference in the task adopted in our study and Vogelsang et al.'s (i.e., general object classification vs face classification) (Jang and Tong, 2021). A recent study using object recognition (Avberšek et al., 2021) reports that the effect of training schedule is consistent with ours. The task difference may be related to the fact that the optimal discriminative features for object recognition are biased toward high frequencies while only low-frequency features are sufficient for good face classification accuracy (Jang and Tong, 2021).

The failure of B2S-Net to recognize blurry images indicates that simply simulating the development of visual acuity during training cannot account for the blur robustness of human vision in general object recognition. However, even after the completion of visual development, we still experience blurred retinal images on a daily basis due to defocus as well as motion blur, and scattering caused by climatic conditions such as rain and fog and by the transmission of translucent objects. In this sense, B+S Net, trained simultaneously

with both blurry and sharp images, can be regarded as reproducing biologically plausible situations to some extent. In addition, the CNNs we tested do not include a mechanism that prevents the forgetting of previously learned representations. B2S-Net may have forgotten the processing for low-frequency components because it was trained only on sharp images in the last 20 epochs. Therefore, B2S-Net may be able to recognize blurred images as well as B+S-Net by adding a mechanism that prevents the model from forgetting the representations tuned for blurred images learned in the early phase of training. Machine learning literature has suggested a few methods to prevent so-called catastrophic forgetting in continual learning (de Melo et al., 2022). One is to protect the weights relevant to the stimuli learned in the early phases of training (Kirkpatrick et al., 2017). This is reminiscent of the critical period of biological neural networks, which strengthens the impact of early childhood experiences on development of the human visual system. Another method is to use the memory of relevant prior information to retrain the network with new information (Aljundi et al., 2019). This mechanism will make the effect of B2S training similar to that of B+S training. With such an additional mechanism against forgetting, B2S-Net may be able to show performance comparable to B+S-Net.

Our results also show that B+S Net has acquired human-level blur robustness but has not acquired human-like global visual processing. The performance test using band-pass filtered images showed that all CNN models, including B+S Net, were not good at utilizing band-limited features while humans retained good accuracy in the mid to high-frequency range. Although the shape bias of the blur-trained models was slightly enhanced, it was not enough to reach the human level. The test using the images with local occlusions revealed that all the models relied primarily on local features, did not utilize the global configuration, and were critically vulnerable to local occlusions. All these results are in stark contrast to human visual processing, which is known to rely more on global configural relationships and shape information and is less sensitive to partial occlusions in object recognition tasks. Therefore, our results indicate that the information processing learned in B+S-Net is still markedly different from that of the human visual system (Geirhos et al., 2021; Baker and Elder, 2022).

Our results reveal that what the networks cannot acquire from blur training is human-recognizable global configuration features present not only in sharp and blurry images but also in high-pass images and texture-shape cue conflict images. Note that the similarity of these classes of images is supported by a finding that the SIN-trained Net shows good recognition for high-pass images as well. In high-pass images, local edge features defined by high-frequency luminance modulations produce global configurations at a scale much larger than a fine-scale edge detector. For detection of these global features, second-order processing such as those modeled by an FRF (filter-rectify-filter) model for human vision [e.g., Graham and Landy (2004)] may be necessary. It seems that object recognition training with sharp and blurred images alone does not provide neural networks with the ability to process second-order features.

According to the comparison of the model architectures, there was no qualitative difference in the effect of blur training. Importantly, we found that VOneNet, which hard-coded the computational processes in the primary visual area (V1) in the front end of AlexNet, did not show improvement in any of the tasks tested in this study. This indicates the limited impact of the initial layer on the frequency tuning at the task performance level and on the mid to high-level information processing related to the shape bias and the configural effect. On the other hand, we also found a few notable differences in the frequency tuning patterns between the architectures. For example, the loss of blur robustness observed in B2S-Net was more prominent in 1000-class AlexNet as well as in 16-class VGG16 and 16-class ResNet50 than in 16-class AlexNet. B+S-Net and B-Net in these architectures were also more narrowly tuned to the blur strength used during training. For the 1000-class AlexNet, the reason for this may be attributed to the fact that the models were exposed to a higher number of images (and thus went through a higher number of weight updates) when using the 1000-class dataset than the 16-class dataset. For the 16-class VGG16 and 16-class ResNet50, differences in model architecture such as increased depth, reduced kernel size, and residual connections (in the case of ResNet) may have resulted in improved learning efficiency, thereby making them more likely to specialize in features that are optimal for the current blur strength. In addition, we also found that VGG16 demonstrated higher accuracy for the band-pass filtered images with high spatial frequency than the other architectures, though we have not yet been able to ascertain why.

In section 4, we analyzed how B+S-Net, which performed similarly to humans in a low spatial frequency image classification task, processed sharp and blurred images. The activity correlation between sharp and blurred images increased in B+S-Net. The results suggest that B+S-Net extracts more common features from sharp and blurred images than S-Net.

The results of zero-shot transfer learning support this view. While the generalization accuracy is not very high, the confusion matrix and the internal activity correlation suggest that B+S training produces a certain degree of common representation between blur and sharp features, which can be used even for unlearned categories.

These results suggest that B+S-Net recognizes sharp or blurred images using common representations, rather than using separate representations. The results also suggest that it is not only linear low-pass filtering in the first layer, but also a series of non-linear processing in the subsequent layers, that produces the common representations. In this respect, we may be able to get useful computational insights into human processing from the analysis of B+S-Net.

Whereas we found B+S training facilitates the development of common processing for sharp (broadband) and blurred (low-pass) images, we found little evidence for B+S training facilitating the development of common processing for low-pass and high-pass images, nor integration of sub-band information. These results suggest that the frequency processing by B+S-Net is critically different from that by the human visual cortex. How can we make the frequency processing more closely resemble that of the human visual system? Several machine learning techniques including data augmentation and contrastive learning may be used to force the network to integrate sub-band information. Note, however, that as a tool to understand human visual computation, it is important that the model training is natural and plausible for the development of the human visual system, like blur training.

In conclusion, training with blurred images provides performance and internal representation comparable to that of humans in recognizing low spatial frequency images. It does narrow,

but only slightly, the gap with the human visual system in terms of global shape information processing and multi-scale frequency information integration.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, the data and codes that support the findings of this study are available at https://github.com/KUCognitiveInformaticsLab/blur-training.

Ethics statement

The studies involving human participants were reviewed and approved by Research Ethics Committee at Graduate School of Informatics, Kyoto University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

SY performed the experiments and data analysis under the supervision of TF and SN, and drafted the first manuscript. All authors contributed to the experimental design, manuscript writing, and approved the submitted manuscript.

Funding

This work was supported by the JSPS Grants-in-Aid for Scientific Research (KAKENHI), Grant Numbers JP20H00603 and JP20H05957.

Conflict of interest

TF and SN were employed by Nippon Telegraph and Telephone Corporation.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1047694/full#supplementary-material

References

Aljundi, R., Caccia, L., Belilovsky, E., Caccia, M., Lin, M., Charlin, L., et al. (2019). "Online continual learning with maximally interfered retrieval," in *Adv. Neural Inf. Process. Syst.* (*NeurIPS)* (Vancouver, BC), 32, 11849–11860.

Avberšek, L. K., Zeman, A., and Open de Beeck, H. (2021). Training for object recognition with increasing spatial frequency: a comparison of deep learning with human vision. *J. Vis.* 21, 14. doi: 10.1167/jov.21.10.14

Baker, N., and Elder, J. H. (2022). Deep learning models fail to capture the configural nature of human shape perception. *iScience* 25, 104913. doi: 10.1016/j.isci.2022.104913

Banks, M. S., and Salapatek, P. (1981). Infant pattern vision: a new approach based on the contrast sensitivity function. *J. Exp. Child Psychol.* 31, 1–45. doi: 10.1016/0022-0965(81)90002-3

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D., DiCarlo, J. J., et al. (2020). "Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations," in *Adv. Neural Inf. Process. Syst.* (*NeurIPS)* (Virtual), Vol. 33, 13073–13087. doi: 10.1101/2020.06.16.154542

de Melo, C. M., Torralba, A., Guibas, L., DiCarlo, J., Chellappa, R., and Hodgins, J. (2022). Next-generation deep learning based on simulators and synthetic data. *Trends Cognit. Sci.* 26, 174–187. doi: 10.1016/j.tics.2021.11.008

Dobson, V., and Teller, D. Y. (1978). Visual acuity in human infants: a review and comparison of behavioral and electrophysiological studies. *Vision Res.* 18, 1469–1483. doi: 10.1016/0042-6989(78)90001-9

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). "Image style transfer using convolutional neural networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* (*CVPR)* (Las Vegas, LA), 2414–2423. doi: 10.1109/CVPR.2016.265

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., et al. (2021). "Partial success in closing the gap between human and machine vision," in *Adv. Neural Inf. Process. Syst.* (*NeurIPS)* (Virtual), 34, 23885–23899.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., Brendel, W., et al. (2019). "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," *Int. Conf. Learn. Represent.* (*ICLR)* (New Orleans, LA).

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., Wichmann, F. A., et al. (2018). "Generalisation in humans and deep neural networks," *Adv. Neural Inf. Process. Syst.* (*NeurIPS)* (Montreal, QC), 31, 7549–7561.

Graham, N., and Landy, M. (2004). "Visual perception of texture," in *The Visual Neurosciences*, eds J. S. Werner, and L. M. Chalupa (Cambridge, MA: MIT Press), 1106-1118. doi: 10.7551/mitpress/7131.003.0084

Grand, R. L., Mondloch, C. J., Maurer, D., and Brent, H. P. (2001). Early visual experience and face processing. *Nature* 410, 890. doi: 10.1038/35073749

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* (*CVPR)* (Las Vegas, LA), 770–778. doi: 10.1109/CVPR.2016.90

Jang, H., and Tong, F. (2021). Convolutional neural networks trained with a developmental sequence of blurry to clear images reveal core differences between face and object processing. *J. Vis.* 21, 6. doi: 10.1167/jov.21.12.6

Jang, H., and Tong, F. (2022). "Lack of experience with blurry visual input may cause cnns to deviate from biological visual systems," in *Abstract of Annual Meeting of Vision Sciences Society (VSS 2022)* (St. Pete Beach, FL). doi: 10.1167/jov.22.14.4324

Jones, J. P., and Palmer, L. A. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.* 58, 1233–1258. doi: 10.1152/jn.1987.58.6.1233

Katzhendler, G., and Weinshall, D. (2019). Potential upside of high initial visual acuity? *Proc. Natl. Acad. Sci. U. S. A.* 116, 18765–18766. doi: 10.1073/pnas.1906400116

Keshvari, S., Fan, X., and Elder, H. J. (2021). "Configural processing in humans and deep convolutional neural networks," in *Abstract of Annual Meeting of Vision Sciences Society (VSS 2021)* (Virtual). doi: 10.1167/jov.21.9.2887

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. U. S. A.* 114, 3521–3526. doi: 10.1073/pnas.1611835114

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," in *Adv. Neural Inf. Process. Syst*. (*NIPS)*, Vol. 25, 1092–1105.

Le Grand, R., Mondloch, C. J., Maurer, D., and Brent, H. P. (2004). Impairment in holistic face processing following early visual deprivation. *Psychol. Sci*. 15, 762–768. doi: 10.1111/j.0956-7976.2004.00753.x

Miller, G. A. (1995). Wordnet: a lexical database for English. *Commun. ACM* 38, 39–41. doi: 10.1145/219717.219748

Nonaka, S., Majima, K., Aoki, S. C., and Kamitani, Y. (2021). Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience* 24, 103013. doi: 10.1016/j.isci.2021.103013

Simonyan, K., and Zisserman, A. (2015). "Very deep convolutional networks for Large-Scale image recognition," in *Int. Conf. Learn. Represent*. (*ICLR)* (San Diego, CA).

van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res*. 9, 2579–2605.

Vaziri-Pashkam, M., Taylor, J., and Xu, Y. (2019). Spatial frequency tolerant visual object representations in the human ventral and dorsal visual processing pathways. *J. Cogn. Neurosci*. 31, 49–63. doi: 10.1162/jocn_a_01335

Vogelsang, L., Gilad-Gutnick, S., Ehrenberg, E., Yonas, A., Diamond, S., Held, R., et al. (2018). Potential downside of high initial visual acuity. *Proc. Natl. Acad. Sci. U. S. A*. 115, 11333–11338. doi: 10.1073/pnas.1800901115

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 10 March 2023
DOI 10.3389/fpsyg.2023.1082557

OPEN ACCESS

EDITED BY
Mary Peterson,
University of Arizona,
United States

REVIEWED BY
Taiki Fukiage,
NTT Communication Science Laboratories,
Japan
Ryan W. Langridge,
University of Manitoba,
Canada

*CORRESPONDENCE
Hiroshige Takeichi
✉ takeichi@a.riken.jp

SPECIALTY SECTION
This article was submitted to
Perception Science,
a section of the journal
Frontiers in Psychology

RECEIVED 28 October 2022
ACCEPTED 22 February 2023
PUBLISHED 10 March 2023

CITATION
Takeichi H, Taniguchi K and
Shigemasu H (2023) Visual and haptic cues in
processing occlusion.
Front. Psychol. 14:1082557.
doi: 10.3389/fpsyg.2023.1082557

Visual and haptic cues in processing occlusion

Hiroshige Takeichi[1,2]*, Keito Taniguchi[3] and Hiroaki Shigemasu[3]

[1]Computational Engineering Applications Unit, Head Office for Information Systems and Cybersecurity (ISC), RIKEN, Wako, Saitama, Japan, [2]Open Systems Information Science Team, Advanced Data Science Project (ADSP), RIKEN Information R&D and Strategy Headquarters (R-IH), RIKEN, Yokohama, Kanagawa, Japan, [3]School of Information, Kochi University of Technology, Kami, Kochi, Japan

Introduction: Although shape is effective in processing occlusion, ambiguities in segmentation can also be addressed using depth discontinuity given visually and haptically. This study elucidates the contribution of visual and haptic cues to depth discontinuity in processing occlusion.

Methods: A virtual reality experiment was conducted with 15 students as participants. Word stimuli were presented on a head-mounted display for recognition. The central part of the words was masked with a virtual ribbon placed at different depths so that the ribbon appeared as an occlusion. The visual depth cue was either present with binocular stereopsis or absent with monocular presentation. The haptic cue was either missing, provided consecutively, or concurrently, by actively tracing a real off-screen bar edge that was positionally aligned with the ribbon in the virtual space. Recognition performance was compared between depth cue conditions.

Results: We found that word recognition was better with the stereoscopic cue but not with the haptic cue, although both cues contributed to greater confidence in depth estimation. The performance was better when the ribbon was at the farther depth plane to appear as a hollow, rather than when it was at the nearer depth plane to cover the word.

Discussion: The results indicate that occlusion is processed in the human brain by visual input only despite the apparent effectiveness of haptic space perception, reflecting a complex set of natural constraints.

KEYWORDS

image segmentation, depth cues, visual pathways, virtual reality, haptic perception

1. Introduction

Occlusion is a typical problem in image processing that involves separating regions that correspond to objects that are apart in the external three-dimensional (3D) space but adjoined in the projected two-dimensional (2D) image because of the proximity of the lines of sight. The occlusion problem comprises two subproblems. First, contours in an image must be segmented. Proper segmentation cannot be obtained by tracking contours in the projection because the contours that are separated in the 3D space may be misleadingly connected in the 2D projection. Second, the segmented contours must be completed to fill in the gaps such that the completed contour is a good estimation of the projected contour without occlusion, that is, occlusion-invariant. Occlusion remains a difficult problem in computer vision (Garcia-Garcia et al., 2017; Albalas et al., 2022). While the input to the second problem of interpolation comprises the output from the first problem of segmentation, the first problem recursively depends on the output from the second problem (Takeichi et al., 1995). For example, if a face is partly occluded and the missing parts are to be completed, one must first know whether there is actually a face without recognizing it. The question is how much of the segmentation can be addressed without committing to interpolation.

The depth relationship appears to be an independent and effective cue for segmentation in general. Segmentation can be significantly changed by inverting the depth relationships between overlapping figures (Nakayama et al., 1990). However, when a stronger cue than the depth is available, depth may not affect the segmentation. For example, the effect of stereoscopic depth is reduced or lost when visual motion is also provided as a segmentation cue (Takeichi, 1999). The problem of segmentation may be solved based on the shape of the contour without occlusion. Completion in occlusion, which is called amodal completion in experimental psychology, may be most parsimoniously predicted by the shape of completed figures (van Lier et al., 1995) in 3D (Tse, 1999, 2017; Kellman et al., 2005). Although dependence on 3D shape apparently supports the iterative computation between segmentation and interpolation, segmentation can still be one-shot if it considers the curvature-based geometric relationship between the 3D shape and its 2D projection (Richards et al., 1987; Elder, 2018). In fact, human perception shows sensitivity to curvature, which is invariant under the projection from 3D to 2D, in processing occlusions (Takeichi, 1995).

Simple segmentation by depth may be performed at an early level in visual processing before the two cortical streams, one for object identity and the other for spatial relationships or interactions (Ungerleider and Mishkin, 1982; Goodale and Milner, 1992), diverged as the ventral and dorsal pathways in the primate visual system (Bakin et al., 2000; O'Herron and von der Heydt, 2013). However, completion that concords with the spatial arrangement of several multipart objects in naturalistic scenes (Tse, 1999, 2017; Kellman et al., 2005) may imply that processing occlusion requires a significant interaction between the two parallel visual pathways. The visual system may not compute the connectedness likelihood only through a simple measure such as simplicity (Buffart and Leeuwenberg, 1981), relatability (Kellman and Shipley, 1991; Kellman et al., 2005) or curvature (Takeichi et al., 1995) but a set of constraints (DiMattina et al., 2012). Tse's (2017) demonstration of amodal completion of fluids or slimy objects further implies that complexity of the constraints may be comparable to "naive" or intuitive physics, such as viscosity, cohesiveness, and specific gravity of fluid and gravitational force. On the one hand, it is natural because processing occlusion is scene analysis. The relevant constraints may span from optics such as the generic view principle (Freeman, 1994; Kitazaki and Shimojo, 1996; Albert, 2001) to laws of mechanics that predict probable and improbable shape, deformation and structure. On the other hand, it also implies that a wide variety of brain areas, particularly those of multisensory integration, may be involved in processing. Material properties that can be related to deformation are estimated in the ventral pathway (Goda et al., 2014), while arrangements and mechanical relationships between several such objects must involve processing in the dorsal pathway. If the perceived property of fluid needs to be integrated with the perceived spatial layout of scattered clusters of such fluid together with the potential occluder to estimate connectedness likelihood in reference to intuitive physics, then the ventral and dorsal pathways must interact as such (c.f. Van Dromme et al., 2016). In addition, whereas it can be hardly tested empirically whether or not the visual input alone is enough for the development of intuitive physics in visual perception, it is also difficult to imagine how concepts such as weight and force develop in visual modality without any reference to tactile or haptic inputs. In fact, perceived occlusion is a purely visual phenomenon because occlusion is defined as interruption of the line of sight. However, it sounds odd if the visual system uses an internal model of intuitive physics to solve the problem only in the visual modality because the intuitive physics itself is likely to be acquired through interaction between visual and tactile or proprioceptive modalities. If processing occlusion is based on intuitive mechanics, then knowledge from previous haptic input, which is the basis of intuitive mechanics, may be used in processing occlusion.

In this study, we investigated the cues that are or are not used in perceptual segmentation and the completion of partially occluded figures. Letters were used as stimulus figures. In the experiment, word recognition performance was compared in the presence and absence of visual and haptic cues to the depth of the occluder. If information provided by haptic input aids in the recognition of partially occluded letters when solving the occlusion or segmentation problem, the presence of haptics may provide information regarding the relationship between the occluded letters and the source of the occlusion. Alternatively, if the effect of haptic input is limited to depth perception, this may imply that haptic inputs have limited roles in recognizing partially occluded letters. The effectiveness of the visual and haptic cues was also evaluated through depth judgments and confidence ratings of the judgments. Confidence was measured because it can be sensitive to potential cue effectiveness (Fairhurst et al., 2018).

2. Methods

A word recognition experiment was performed to examine perceptual completion using virtual reality (VR). Word recognition performance was assessed when the central part of the word stimulus was masked by a horizontal virtual ribbon. The virtual reality experiment was conducted with 15 students as participants. The participants had to fill in the missing central part by connecting the top and bottom parts that remained in the visual stimulus to recognize the words. The edges of the ribbon were implicit and invisible, as if the ribbon were camouflaged to have the same lightness, color, and texture as the background. Cues to the depth of the ribbon were provided visually through random dots scattered over the surface of the ribbon, haptically by active tracing of the edge of the ribbon, or both. This unnatural occlusion was simulated to reduce visual cues. Visual input is provided as a stream of two-dimensional arrays of pixels, i.e., images, while haptic input is provided as a time-series of points by scanning the target over time. Because it is difficult to control two-dimensional haptic exploration to give comparable inputs in both visual and haptic modalities, the haptic input was limited to the edges. The visual input was thus similarly limited to the edges with the obscure occluder.

2.1. Ethics statement

Data collection and processing were performed in accordance with the principles of the Declaration of Helsinki. The protocols of the human experiments in this study were approved prior to initiation by the institutional review board of Kochi University of Technology (138-C2) and Wako Third Ethical Committee of RIKEN (Wako3 2020-27).

2.2. Participants

A total of 15 volunteers (13 men and 2 women; 22.06 years old with a standard deviation of 0.92 years) participated in the experiment. They were students at Kochi University of Technology or their affiliates. All participants had normal or corrected-to-normal vision and normal binocular stereopsis, which was confirmed with the experimental setup before the experiment, and were native Japanese readers. Written informed consent was obtained from each participant before participation.

2.3. Stimulus and apparatus

A visual-haptic multimodal stimulus was presented as a type of mixed reality. The visual stimulus (Figure 1) comprising words for visual recognition and a horizontal virtual ribbon that masked the central part of the word was presented on a head-mounted display (Oculus Rift CV1). The spatial resolution of the head-mounted display was 1,080 by 1,200 per eye, and the diagonal field of view was approximately 110 degrees. The refresh rate was 90 Hz. The virtual ribbon simulated an occlusion when it appeared to be above and covering the central part of the word or a hollow when it appeared to be farther in depth at the central part of the word. The word stimuli were five-letter Japanese words written in katakana characters, which are alphabet-like phonograms in Japanese. The words were obtained from an open database for teaching Japanese as a second language and, therefore, were quite commonly used by native readers. A total of 337 five-letter words were extracted from the database, excluding words containing one or more small characters to indicate the palatalized "y" sound or double consonants in Japanese. The characters were displayed using public-domain font-type FAMania for the ease of camouflage, which mimics low-resolution (7×7 pixels) characters that were used on gaming PCs during the 1980s. The ribbon was positioned such that its edges naturally coincided with pixel boundaries, that is, between the first and the second rows and between the sixth and the seventh rows of the pixel matrix. If a font type with a higher resolution was used, placing the virtual ribbon would inevitably introduce conspicuous linearly aligned terminators, which could be a strong cue to depth discontinuity, such as abutting gratings that induce perception of illusory contours (Soriano et al., 1996). The background was gray, with scattered black random dots of 15% density. The random dots had a binocular disparity of 22.3 min in the conditions with binocular cues to depth.

2.4. Design and procedure

In each trial, a word was presented with the virtual ribbon. The task was to verbally report the word recognized after stimulus presentation. There were 12 combinations of two levels of depth, two types of visual cues, and three types of haptic cues. The depth of the

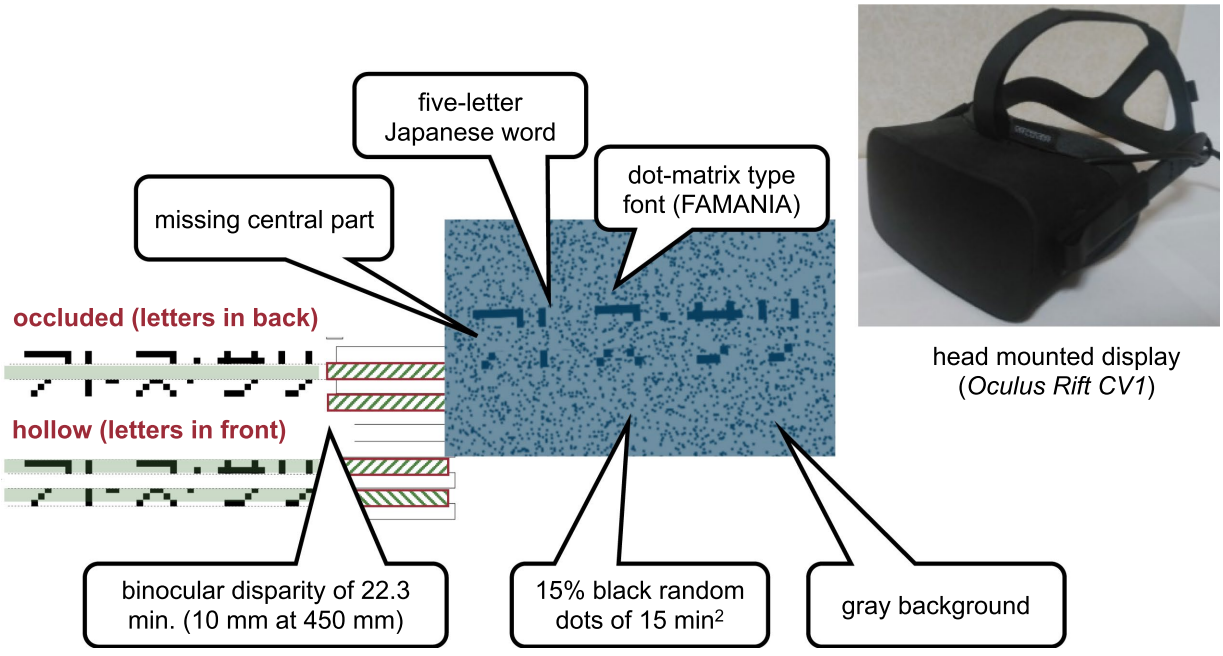


FIGURE 1
Visual stimulus. The haptic stimulus (Figure 2) was a wooden board. Four horizontal square-wave bumps were created on the board by gluing square columns in parallel onto the surface. Multiple bumps were made to change the physical position of the edge across trials to eliminate artifactual cues and interferences from a nonvarying stimulation. The edges of the square columns were aligned with the edges of the virtual ribbon in the VR space. The participants actively traced one of the edges of one of the columns using the tip of the index finger of their right hands to receive haptic input (Figure 3). An Oculus Touch controller was used for positional alignment between vision and haptics. The participants held the Oculus Touch controller while tracing the specified edge. The position of the fingertip was visually indicated by a blue virtual ball that moved in real-time synchrony with the motion of Oculus Touch in the head-mounted display. The VR system was implemented using Unity 2017.4.15f1. The latency was within 25ms after the movement onset and then within 5ms on average (Warburton et al., 2022). The participant sat on a chair and was confronted with a wooden board on a desk 45cm in front of them. A chinrest was used to minimize head movement.

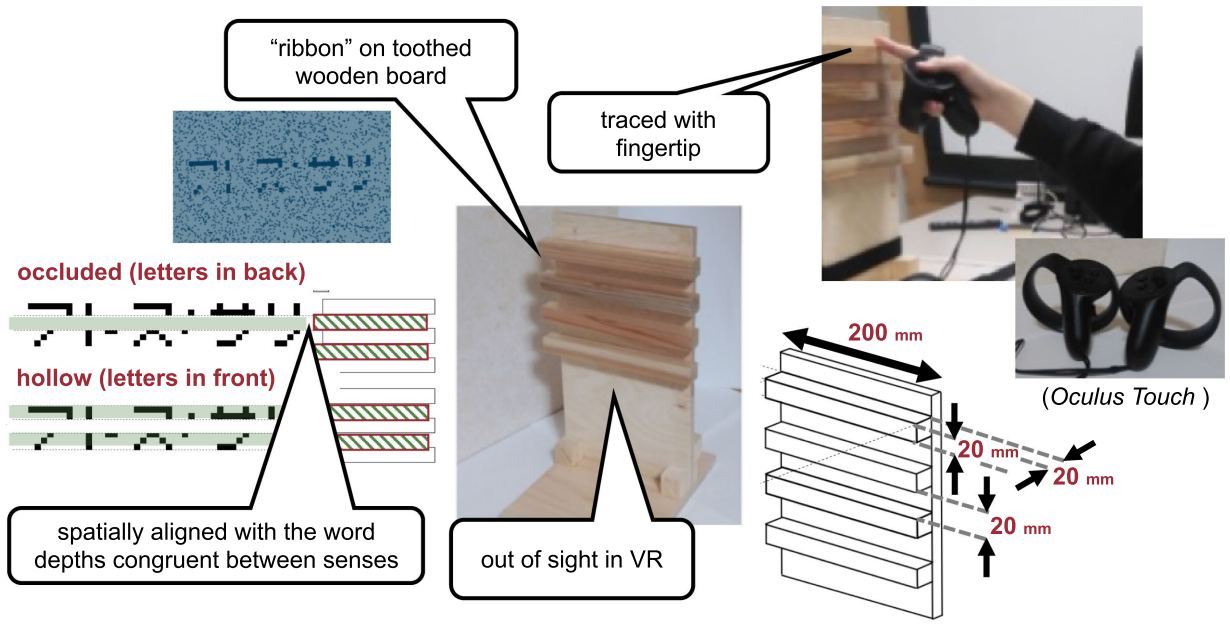


FIGURE 2
Haptic stimulus.

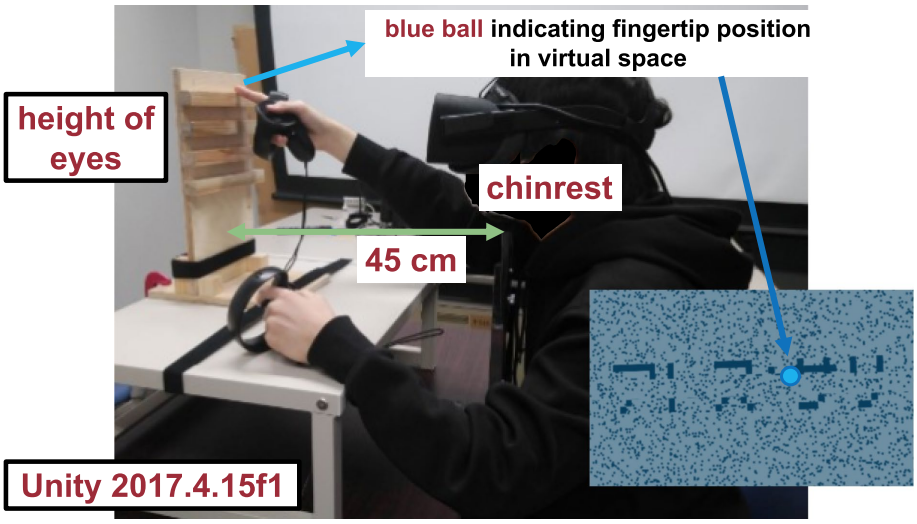


FIGURE 3
Experimental setup.

word was either in the background of the masking virtual ribbon that appeared to be an occlusion or in front of the virtual ribbon that appeared to be hollow. The visual cue was either present as a horizontal binocular disparity of the dots in the area covered by the virtual ribbon or absent by monocular presentation with a blank screen in the other eye. The haptic cue was either absent, consecutive, or concurrent. When the haptic cue was absent, the visual stimulus (i.e., the word and the virtual ribbon) was presented for 10 s, with the ribbon as an occlusion or a hollow, and with or without binocular disparity, followed by a uniform gray screen for another 10 s. The participant was instructed to report the recognized word verbally during the latter 10 s. When the haptic cue was concurrent, the participants were additionally instructed to actively and concurrently trace one of the edges of the ribbon, but not the word, with the fingertip during the former 10 s period simultaneously with visual presentation. When the haptic cue was consecutive, a pair of red balls was first visually presented against a uniform gray background before the visual presentation of the word. The red balls indicate the positions of the ends of the edge in the VR space. The participants were instructed to actively trace the camouflaged edge with the fingertip for the first 5 s and were instructed to take the finger off the board during the following 10 s period for visual presentation. Thus, the edge was not

traced when the haptic cue was absent, was traced simultaneously with the visual presentation when the haptic cue was concurrent, and was traced only during the first 5 s without visual stimulus but not during the following 10 s with the visual stimulus when the haptic cue was consecutive.

Each condition was presented in a block of five trials, in which five different words were presented. The 12 conditions were presented in a set of 12 blocks in randomized order. Each participant performed four sets on 2 days. Therefore, there were 240 trials for each participant. Each condition was repeated 20 times in four sets of five blocked trials. The same set of words was used for all participants. Different words were used in different trials so that the same word was used only once for each participant. Different words were randomly assigned to various conditions across the participants. The eye stimulated for the monocular presentation alternated between successive blocks of the monocular conditions. The participants traced the upper or lower edge of the ribbon with the palm facing downward or upward, respectively, in the two halves of the blocks in the occlusion condition. Similarly, they traced the upper or lower edge of the ribbon with the palm facing upward or downward, respectively, in the two halves of the blocks in the hollow condition. The edge to be traced and direction of the palm were varied to eliminate the potential association between the hand shape and depth as unintended cues. The potential association could provide information regarding the depth artifactually and, therefore, could contaminate the results. The hand shape was specified by instructions for each block and alternated between successive blocks of consecutive and concurrent haptic cue conditions. The edge to be traced varied randomly across trials among the three alternatives with different heights to minimize the potential effect of position in the visual field or peri-personal space. The participants also had to report the perceived depth with confidence on a seven-point scale ranging from −3 to 3 at the end of each block. The confidence as well as the perceived depth was measured in order to identify the extent to which there was an effect of haptic input: whether the haptic input does not influence perceived depth at all or it does influence perceived depth but not recognition. The participants could take breaks at will at any block interval.

The participants' head position was continuously monitored by a computer, and a trial was aborted if the computer detected a displacement larger than 30 mm away from the original position. The same condition was then conducted later in an additional trial to fill in the missing observation. The participants were informed of the constraint during the instruction. The threshold of 30 mm was empirically determined such that any intentional head movement was detected, which resulted in the participant's voluntary immobilization and few aborted trials. Therefore, although there could be some influence of head movement and motion parallax because the stimuli were not presented statically regardless of the participants' head position, our method must have led to the most natural and cost-effective suppression of potential artifacts of head motion and motion parallax.

2.5. Statistics

The rate of correct word recognition was calculated for each of the 12 conditions for each of the 15 participants across 20 repetitions. The stimulus word was considered correctly recognized (score = 1/20) if the response fully matched the five-letter stimulus word and not (score = 0/20) otherwise. The rate of correct depth perception was calculated but not evaluated because it was saturated by the ceiling. The mean confidence rating was calculated and evaluated as the mean of the absolute value of the perceived depth on the seven-point scale for each of the 12 conditions for each of the 15 participants across the four repetitions. A three-way analysis of variance was performed with repeated measurements using Anova-kun 4.8.5 (Iseki, 2020) on R 4.0.2 (R Core Team, 2018) for each of the recognition rate and the mean depth confidence rating. The factors were depth, visual cues, and haptic cues: Depth was either letters-in-front or letters-in-background; the visual cue was either present through binocular stereopsis or absent through monocular presentation; and the haptic cue was absent, consecutive, or concurrent. Multiple comparisons were performed using Shaffer's method for the corrected alpha of 0.05 for individual ANOVA of the two different indices: recognition rate and depth rating. Violations to sphericity were tested using Mendoza's multisample sphericity test. In addition, to examine the effects of deviation from the normal distribution regarding the word recognition rate, a mixed model analysis was performed assuming a binomial in place of normal distribution after transformation by a logistic function.

The general linear mixed models included fixed factors of depth, visual cue, haptic cue and all their second- and third-order interactions and random factors of participant and position of the three-alternative traced edge. The analyses with general linear mixed models were performed using MATLAB R2021a or later.

2.6. Additional analyses

It is worth noting how much of the recognition performance measured in this task reflected the success of the amodal completion process. Some characters might be too hard or even impossible to identify due to occlusion of a critical part. In such cases, amodal completion would not necessarily contribute to correct answers. Thus, the same general linear mixed model was tested with the data after removing letters that were deemed too difficult to complete, as follows. First, the mean correct response rate was calculated for each of the 67 characters regardless of the condition or the participant. Second, the characters with relatively poor correct responses, namely, <0.8, were identified. Finally, the scores were recalculated based only on the remaining "easy" characters that could be recognized in more than 80% of the cases to be examined.

Responses with long latencies up to 10 s were allowed to accommodate the time that the participants needed, especially in the concurrent haptic cue condition, in which they performed a dual task. However, because responses with longer latencies are generally based more on cognitive processes, an analysis that is limited to data with shorter latencies may focus more on perceptual processes. Thus, word recognition was also evaluated when the data were limited to trials with response times shorter than the overall mean response time.

Each word was presented only once for each participant. However, potential differences between characters need further consideration. Because the order of the words was different between participants, different characters appeared at different positions along the progression of the experiment. Therefore, potential character-specific learning could have resulted in either a spurious bias, i.e., false

positives, or extraneous variability that may contribute to error variance, i.e., false negatives. To rule out these potential artifacts, a measure of character-specific learning was constructed. For each of the 5 characters in the word stimulus, the frequency of appearance in all preceding stimuli was enumerated and summed up to be a "familiarity" index of the word for each trial of each participant. The familiarity index was added to the predictors in the analysis of the potential effects of perceptual learning. The last two analyses additionally included a random factor of interaction between position of the traced edge and trial number.

3. Results

Figure 4 shows the results of word recognition. The correct word-report rate is shown for each of the 12 conditions. Recognition was better in the visual cue present (*Vs*+) condition than in the visual cue absent (*Vs*-) condition [$F(1, 14) = 11.51$, $p < 0.01$, $\eta^2 = 0.027$] and in the letters-in-front (Fr) condition than in the letters-in-background (Bk) condition [$F(1, 14) = 5.13$, $p < 0.05$, $\eta^2 = 0.008$]. The type of haptic cue (Hp+&, Hp+, and Hp−) did not have a significant effect [$F(2, 28) = 0.88$, $p = 0.45$, $\eta^2 = 0.003$]. None of the interactions were significant [Haptic-cue and depth: $F(2, 28) = 0.18$, $p = 0.83$, $\eta^2 = 0.003$; Haptic-cue and Visual-cue: $F(2, 28) = 0.53$, $p = 0.59$, $\eta^2 = 0.001$; Visual-cue and depth: $F(1, 14) = 0.27$, $p = 0.60$, $\eta^2 = 0.000$; Visual-cue, Haptic-cue and depth: $F(2, 28) = 1.12$, $p = 0.33$, $\eta^2 = 0.003$]. No violations to the sphericity test were found. The results were essentially the same when a linear mixed model was evaluated with a binomial distribution and the logistic link function. The effect of depth was $F(1, 168) = 4.961$, $p = 0.02725$, Cohen's $f^2 = 0.03018$ and $F(1, 168) = 4.992$, $p = 0.02679$, $f^2 = 0.02707$ assuming a normal distribution and a binomial distribution, respectively. The effect of the visual cue was $F(1,168) = 15.82$, $p = 0.00001033$, $f^2 = 0.09624$ and $F(1,168) = 15.94$, $p = 0.00009746$, $f^2 = 0.07766$ assuming a normal distribution and a binomial distribution, respectively.

Whereas the criterion of 80% correct recognition was arbitrary, it separated 44 "easy" characters from 23 "difficult" characters, which were "クグシゼソゾタダチヅネハパヒフプベペメヤユヨラ." As a result of the analysis of the score that was recalculated only on the 44 easy characters with general linear mixed models, the effects of depth and visual cue remained the only significant factors [$F(1,168) = 4.785$, $p = 0.03009$, $f^2 = 0.02909$; $F(1,168) = 14.40$, $p = 0.0002061$, $f^2 = 0.08755$]. The effect of the haptic cue reached closer to the significance level [$F(2,168) = 2.065$, $p = 0.1301$, $f^2 = 0.02510$]. As a result of the general linear mixed model that included the familiarity index as a fixed factor, the effect of the familiarity index of character-specific learning was highly significant [$F(1,3576) = 48.84$, $p < 0.00001$, $f^2 = 0.007416$]. The only significant interaction with the familiarity index was with the visual cue [$F(1,3576) = 5.015$, $p = 0.02519$, $f^2 = 0.001344$].

There were fewer errors in depth judgments with the visual cue (21.11% of the trials) than in depth judgments without the visual cue (49.72%). The confidence in the depth rating is shown for each of the 12 conditions in Figure 5. The rating was more confident in the visual cue present condition than in the visual cue absent condition [$F(1, 14) = 60.60$, $p < 0.001$, $\eta^2 = 0.575$]. There was also an effect of the type of haptic cue [$F(2, 28) = 4.84$, $p < 0.05$, $\eta^2 = 0.014$], and confidence was larger in the consecutive-haptic-cue condition than in the haptic-cue-absent and concurrent-haptic-cue conditions after corrections for multiple comparisons (Shaffer method, $p < 0.05$). Furthermore, the interaction between visual and haptic cues was significant [$F(2, 28) = 6.17$, $p < 0.01$, $\eta^2 = 0.006$]. When the visual cue

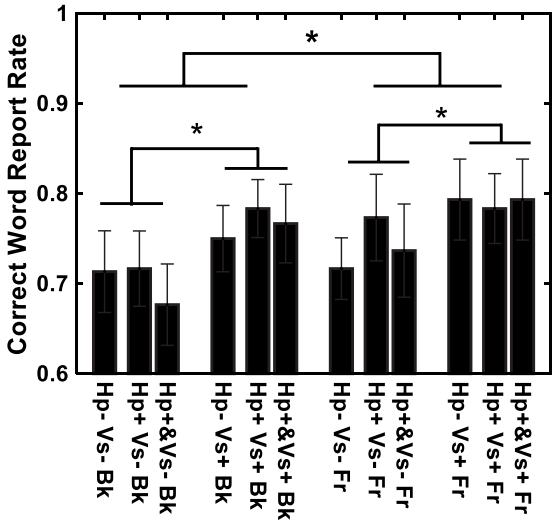


FIGURE 4
Recognition performance for individual conditions. *Vs*+: visual cue present with binocular stereopsis. *Vs*−: visual cue absent with monocular presentation. Hp+&: haptic cue to visual presentation. Hp+: haptic cue subsequently to visual presentation. Hp−: no haptic cue. Fr: virtual ribbon in front of the word. Bk: virtual ribbon as a hollow the word in the background. *: difference statistically significant.

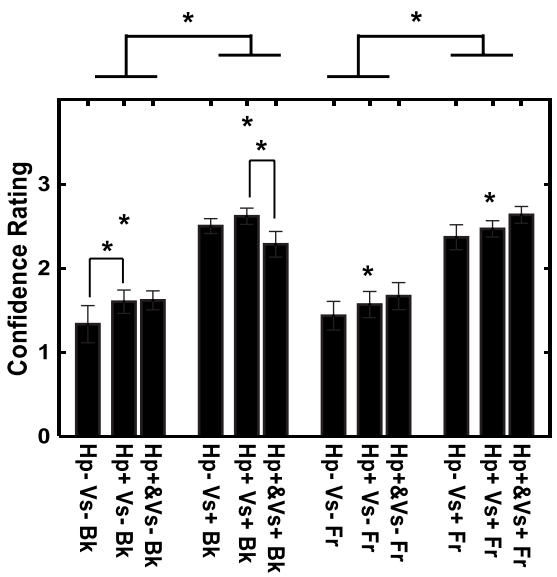


FIGURE 5
Confidence in depth perception for individual conditions. *Vs*+: visual cue present with binocular stereopsis. *Vs*−: visual cue absent with monocular presentation. Hp+&: haptic cue with visual presentation. Hp+: haptic cue subsequently to visual presentation. Hp−: no haptic cue. Fr: virtual ribbon in front of the word. Bk: virtual ribbon as a hollow in the word in the background. *: difference statistically significant.

was absent, the participants had more confidence in the consecutive-haptic-cue condition than in the haptic-cue absent condition. When the visual cue was present, the consecutive-haptic-cue condition was better than the concurrent-haptic-cue condition.

If the data were limited to the trials with response times shorter than the overall mean response time, which was 4.7 s, then none of the effects were significant, except that the effect of the haptic cue was close to the significance level [F (2,1438) = 2.926, p = 0.05391, f^2 = 0.004102]. The response time was longer in the concurrent haptic cue condition than in the other two haptic cue conditions, regardless of the depth and visual cue conditions (Figure 6, F (2,2517) = 162.8, p < 0.00001, f^2 = 0.1304).

4. Discussion

Word recognition was better in the presence of binocular visual cues to depth but showed an atypical pattern of asymmetry in terms of the sign of the depth relationship; recognition performance was better in the letter-in-front condition than in the letter-in-background condition. Furthermore, the performance was not better in the presence of haptic cues to depth, although the participants reported higher confidence in depth perception with the presence of haptic cues. One interpretation of the atypical effect of the visual cue is that it enhanced but did not alter the segmentation based on the shape of the nonoccluded parts of the word stimuli. Another interpretation is that the occluder was made of sparse dots; therefore, there was less difference between the occluding surface and camouflaging texture in the background. One may be tempted to argue that the absence of a significant interaction between visual cue (*Vs*+ versus *Vs*−) and depth (Fr versus Bk) suggests an effect of binocular summation but not that of binocular stereopsis as the source of the effect of visual cue. Whereas better performance with the visual cue (*Vs*+ > *Vs*−) may indicate the effect of binocular summation, the pattern of the results is also consistent with the effect of binocular stereopsis. Namely, the difference between the depth conditions is larger with the visual cue (Hp−*Vs*+Bk versus Hp−*Vs*+Fr) than without the visual cue (Hp−*Vs*−Bk versus Hp−*Vs*−Fr) in the absence of interference from the haptic cues. The letter-in-front condition led to slightly better performance, probably because nearer stimuli are perceptually more salient. The apparent interference between the visual and concurrent haptic cues might have stemmed from the participant's limited cognitive resources, such as attention or technical imperfections in the alignment between the visual and haptic presentations. Overall, the results suggest that perceptual completion is only visually depth-based and that haptic cues may not enhance segmentation.

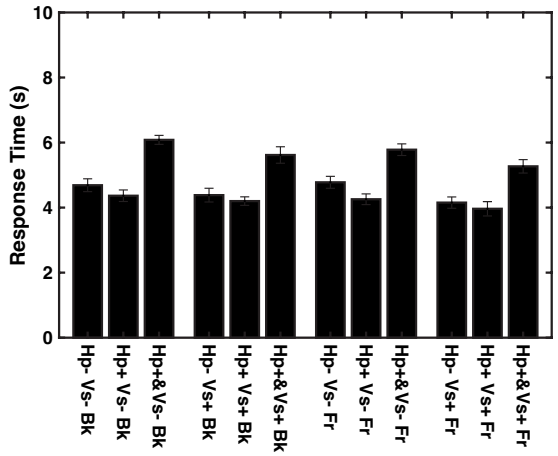


FIGURE 6
Response times for individual conditions. *Vs*+: visual cue present with binocular stereopsis. *Vs*-: visual cue absent with monocular presentation. Hp+&: haptic cue with visual presentation. Hp+: haptic cue subsequently to visual presentation. Hp−: no haptic cue. Fr: virtual ribbon in front of the word. Bk: virtual ribbon as a hollow in the word in the background.

4.1. Potential effects of cognitive factors and perceptual learning

The results were essentially the same when the scores were recalculated based only on "easy" characters that could be recognized in more than 80% of the cases. Some of the difficult characters are likely to be low-frequency characters in Japanese. For example, three characters "パプペ" among the 23 difficult characters have small circles at the top right corner to indicate voiceless "p" sound that mostly appears in loanwords.

There were no significant effects when the valid responses were limited to fast responses. However, it may not necessarily indicate that the present results merely reflect cognitive factors. The response times in the concurrent haptic cue condition were longer than those in the other two conditions. Thus, the division of data by response time effectively separated the data between conditions, thereby eliminating the differences by available cues. The longer response times in the concurrent haptic cue condition may be related to parallel processing of doubled information in a dual task rather than more top-down factors, as supported by the confidence in perceived depth. The depth seemed perceived more confidently with one or more cues than without cues but not in the concurrent haptic cue condition. Concurrent haptic processing seems to have interfered with visual processing when both cues were provided.

The result of the general linear mixed model with the familiarity index showed a highly significant effect of character-specific learning. However, the only significant interaction with the familiarity index was that with the visual cue, which suggested a decreasing effect of the visual cue as character-specific learning proceeded. It does not seem to have altered the potential effects of haptic cues in either way. Therefore, while character-specific learning took place, it does not seem to have altered the results.

4.2. Letter specificity

One potential reason why completion was shape-dominant rather than depth-dominant is that the figures to be completed were letters. Letters differ from other more general objects for computational and biological reasons. They are computationally 2D because they are not projections of 3D objects. Letter recognition is biologically exceptional in its automaticity, as demonstrated by the Stroop effect (Stroop, 1935) and specific and localized neural responses in several measurement modalities (Fujimaki et al., 1999; Cohen et al., 2000; Maurer et al.,

2005). If this is the case, then different results could be obtained with various types of stimuli.

However, the absence of an effect of haptic input may still be related to occlusion being an inherently visual phenomenon. Processing occlusion may be only in the visual modality because occlusion is represented in the egocentric visual space that is explored with line of sight and not in the haptic space that is explored with points of touch. There is no occlusion in haptics because occlusion interrupts the line of sight, and there is no line of sight for haptics. Simultaneously, it has been reported that perception becomes haptic or tactile dominant when somatosensory input is more reliable than visual input (Ernst and Banks, 2002). If this is the case, haptic input might have some effects when the occlusion cannot be processed reliably by the generic view principle (Albert, 2001) because the vantage point is accidental and the visual input is severely limited.

4.3. Cortical site of processing occlusion

As the cortical area for perceptual completion of partially occluded letters is suggested by the results, area LO-1 is the best candidate for the neural correlate of the computation characterized through the experiment. Although the homology between human and nonhuman primate brains is not straightforward, the monkey counterpart of the human lateral occipital complex LOC (Grill-Spector et al., 2001) or the LO area in the ventral processing of object recognition is most likely the inferotemporal cortex, wherein the computation of occlusion-invariant representation is observed (Namima and Pasupathy, 2021). The lateral occipital area LO in the human brain can be divided into two subregions (Wandell et al., 2007), and subregion LO-2 shows greater shape selectivity than LO-1 (Silson et al., 2013; Vernon et al., 2016). As subregion LO-1 is adjacent to the human homolog of area V4, area LO-2 likely overlaps with area LOvt, which shows responsiveness to haptic input (Amedi et al., 2002; Monaco et al., 2017). Thus, the properties of the lower subregion LO-1 match shape processing without haptic input, suggesting that it is a good candidate for the area responsible for the perceptual completion of letters. Coactivation has also been reported between the LO area and the areas related to processing letters and words in the ventral occipitotemporal areas along the fusiform gyrus in the left hemisphere (Agrawal et al., 2020). LO-1 may be a good candidate considering the interaction between the dorsal and ventral pathways because LO-1 is closer to V3A/B, which belongs to the dorsal pathway, than LO-2. In fact, the vertical occipital fasciculus (VOF), or the fiber that connects the area LO and other ventral and lateral visual areas and the area V3A/B, has been identified in the human brain using a combination of fMRI, diffusion MRI and fiber tractgraphy (Takemura et al., 2017). Connections rather than individual areas may be more appropriate neural correlates of complex computations that are commonly found in human visual perception.

The present result may suggest unimodal three-dimensional representation as a precursor to multisensory three-dimensional representation. Constructing representation of the occluded part must be distinguished from recognizing the partly occluded or partly missing figures or objects. Computationally, recognition can be achieved without explicit representation, whereas explicit representation should help recognition. In other words, whereas explicit representation or completion is sufficient for recognition, it is not necessary. There must be some reason or computational benefit of actively constructing explicit representation or actively assuming presence rather than passively ignoring the absence of input from the invisible part. One such potential benefit is to predict the invisible part for future bodily interaction or tangibility. This may correspond to a bifurcation of the flow of information to the pathways for action and recognition (Goodale and Milner, 1992).

5. Conclusion

A VR experiment was conducted to investigate the effects of visual and haptic cues on recognizing partially occluded letters. Although visual cues enhanced letter recognition, enhancement was not specific to the sign of the depth relationships that are typical for occlusion. Haptic cues had no effect on recognition. The results suggest that LO-1 is the most likely cortical locus of the core for processing occlusion, although it must be examined in future studies whether the results are specific to letters or whether visual input dominates even when visual input is singular and haptic input provides comparatively more reliable or useful information. In either case, the present results illustrate how biological processing mirrors a complex set of natural constraints in processing occlusion.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by Institutional Review Board of Koch University of Technology (138-C2) and Wako Third Ethical Committee of RIKEN (Wako3 2020-27). The patients/participants provided their written informed consent to participate in this study.

Author contributions

HT and HS had a role in the conception, design, and interpretation of data. KT and HS had a role in the acquisition and analysis. HT and KT had a role in drafting the work. HS had a role in revising it critically. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by KAKENHI 20K03500 from the Japan Society for Promotion of Science.

Acknowledgments

The authors are grateful to Valter Ciocca for his advice on statistics.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1082557/full#supplementary-material

References

Agrawal, A., Hari, K., and Arun, S. P. (2020). A compositional neural code in high-level visual cortex can explain jumbled word reading. *elife* 9:e54846. doi: 10.7554/eLife.54846

Albalas, F., Alzu'bi, A., Alguzo, A., al-Hadhrami, T., and Othman, A. (2022). Learning discriminant spatial features with deep graph-based convolutions for occluded face detection. *IEEE Access* 10, 35162–35171. doi: 10.1109/ACCESS.2022.3163565

Albert, M. K. (2001). Surface perception and the generic view principle. *Trends Cogn. Sci.* 5, 197–203. doi: 10.1016/s1364-6613(00)01643-0

Amedi, A., Jacobson, G., Hendler, T., Malach, R., and Zohary, E. (2002). Convergence of visual and tactile shape processing in the human lateral occipital complex. *Cereb. Cortex* 12, 1202–1212. doi: 10.1093/cercor/12.11.1202

Bakin, J. S., Nakayama, K., and Gilbert, C. D. (2000). Visual responses in monkey areas V1 and V2 to three-dimensional surface configurations. *J. Neurosci.* 20, 8188–8198. doi: 10.1523/JNEUROSCI.20-21-08188.2000

Buffart, H., and Leeuwenberg, E. (1981). Coding theory of visual pattern completion. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 241–274. doi: 10.1037//0096-1523.7.2.241

Cohen, L., Dehaene, S., Naccache, L., Lehéricy, S., Dehaene-Lambertz, G., Hénaff, M. A., et al. (2000). The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain* 123, 291–307. doi: 10.1093/brain/123.2.291

DiMattina, C., Fox, S. A., and Lewicki, M. S. (2012). Detecting natural occlusion boundaries using local cues. *J. Vis.* 12:15. doi: 10.1167/12.13.15

Elder, J. H. (2018). Shape from contour: computation and representation. *Annu. Rev. Vis. Sci.* 4, 423–450. doi: 10.1146/annurev-vision-091517-034110

Ernst, M. O., and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 429–433. doi: 10.1038/415429a

Fairhurst, M. T., Travers, E., Hayward, V., and Deroy, O. (2018). Confidence is higher in touch than in vision in cases of perceptual ambiguity. *Sci. Rep.* 8:15604. doi: 10.1038/s41598-018-34052-z

Freeman, W. T. (1994). The generic viewpoint assumption in a framework for visual perception. *Nature* 368, 542–545. doi: 10.1038/368542a0

Fujimaki, N., Miyauchi, S., Pütz, B., Sasaki, Y., Takino, R., Sakai, K., et al. (1999). Functional magnetic resonance imaging of neural activity related to orthographic, phonological, and lexico-semantic judgments of visually presented characters and words. *Hum. Brain Mapp.* 8, 44–59. doi: 10.1002/(SICI)1097-0193(1999)8:1<44::AID-HBM4>3.0.CO;2-#

Garcia-Garcia, A., Garcia-Rodriguez, J., Orts-Escolano, S., Oprea, S., Gomez-Donoso, F., and Cazorla, M. (2017). A study of the effect of noise and occlusion on the accuracy of convolutional neural networks applied to 3D object recognition. *Comput. Vis. Image Underst.* 164, 124–134. doi: 10.1016/j.cviu.2017.06.006

Goda, N., Tachibana, A., Okazawa, G., and Komatsu, H. (2014). Representation of the material properties of objects in the visual cortex of nonhuman primates. *J. Neurosci.* 34, 2660–2673. doi: 10.1523/JNEUROSCI.2593-13.2014

Goodale, M. A., and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends Neurosci.* 15, 20–25. doi: 10.1016/0166-2236(92)90344-8

Grill-Spector, K., Kourtzi, Z., and Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vis. Res.* 41, 1409–1422. doi: 10.1016/s0042-6989(01)00073-6

Iseki, R. (2020). Available at: http://riseki.php.xdomain.jp/index.php?ANOVA%E5%90%9B

Kellman, P. J., Garrigan, P., and Shipley, T. F. (2005). Object interpolation in three dimensions. *Psychol. Rev.* 112, 586–609. doi: 10.1037/0033-295X.112.3.586

Kellman, P. J., and Shipley, T. F. (1991). A theory of visual interpolation in object perception. *Cogn. Psychol.* 23, 141–221. doi: 10.1016/0010-0285(91)90009-d

Kitazaki, M., and Shimojo, S. (1996). 'Generic-view principle' for three-dimensional-motion perception: optics and inverse optics of a moving straight bar. *Perception* 25, 797–814. doi: 10.1068/p250797

Maurer, U., Brandeis, D., and McCandliss, B. D. (2005). Fast, visual specialization for reading in English revealed by the topography of the N170 ERP response. *Behav. Brain Funct.* 1:13. doi: 10.1186/1744-9081-1-13

Monaco, S., Gallivan, J. P., Figley, T. D., Singhal, A., and Culham, J. C. (2017). Recruitment of Foveal Retinotopic cortex during haptic exploration of shapes and actions in the dark. *J. Neurosci.* 37, 11572–11591. doi: 10.1523/JNEUROSCI.2428-16.2017

Nakayama, K., Shimojo, S., and Ramachandran, V. S. (1990). Transparency: relation to depth, subjective contours, luminance, and neon color spreading. *Perception* 19, 497–513. doi: 10.1068/p190497

Namima, T., and Pasupathy, A. (2021). Encoding of partially occluded and occluding objects in primate inferior temporal cortex. *J. Neurosci.* 41, 5652–5666. doi: 10.1523/JNEUROSCI.2992-20.2021

O'Herron, P., and von der Heydt, R. (2013). Remapping of border ownership in the visual cortex. *J. Neurosci.* 33, 1964–1974. doi: 10.1523/JNEUROSCI.2797-12.2013

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria

Richards, W. A., Koenderink, J. J., and Hoffman, D. D. (1987). Inferring three-dimensional shapes from two-dimensional silhouettes. *J. Opt. Soc. Am. A* 4, 1168–1175. doi: 10.1364/JOSAA.4.001168

Silson, E. H., McKeefry, D. J., Rodgers, J., Gouws, A. D., Hymers, M., and Morland, A. B. (2013). Specialized and independent processing of orientation and shape in visual field maps LO1 and LO2. *Nat. Neurosci.* 16, 267–269. doi: 10.1038/nn.3327

Soriano, M., Spillmann, L., and Bach, M. (1996). The abutting grating illusion. *Vis. Res.* 36, 109–116. doi: 10.1016/0042-6989(95)00107-b

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *J. Exp. Psychol.* 18, 643–662. doi: 10.1037/h0054651

Takeichi, H. (1995). The effect of curvature on visual interpolation. *Perception* 24, 1011–1020. doi: 10.1068/p241011

Takeichi, H. (1999). The effects of stereoscopic depth on completion. *Percept. Psychophys.* 61, 144–150. doi: 10.3758/BF03211955

Takeichi, H., Nakazawa, H., Murakami, I., and Shimojo, S. (1995). The theory of the curvature-constraint line for amodal completion. *Perception* 24, 373–389. doi: 10.1068/p240373

Takemura, H., Pestilli, F., Weiner, K. S., Keliris, G. A., Landi, S. M., Sliwa, J., et al. (2017). Occipital white matter tracts in human and macaque. *Cereb. Cortex* 27, 3346–3359. doi: 10.1093/cercor/bhx070

Tse, P. U. (1999). Volume completion. *Cogn. Psychol.* 39, 37–68. doi: 10.1006/cogp.1999.0715

Tse, P. U. (2017). Dynamic volume completion and deformation. *Perception* 8:204166951774036. doi: 10.1177/2041669517740368

Ungerleider, L. G., and Mishkin, M. (1982). "Two cortical visual systems" in *Ch. 18, Analysis of Visual Behavior*. eds. D. J. Ingle, M. A. Goodale and R. J. W. Mansfield (Cambridge: MIT Press)

Van Dromme, I. C., Premereur, E., Verhoef, B. E., Vanduffel, W., and Janssen, P. (2016). Posterior parietal cortex drives inferotemporal activations during three-dimensional object vision. *PLoS Biol.* 14:e1002445. doi: 10.1371/journal.pbio.1002445

van Lier, R. J., van der Helm, P. A., and Leeuwenberg, E. L. (1995). Competing global and local completions in visual occlusion. *J. Exp. Psychol. Hum. Percept. Perform.* 21, 571–583. doi: 10.1037//0096-1523.21.3.571

Vernon, R. J., Gouws, A. D., Lawrence, S. J., Wade, A. R., and Morland, A. B. (2016). Multivariate patterns in the human object-processing pathway reveal a shift from retinotopic to shape curvature representations in lateral occipital areas, LO-1 and LO-2. *J. Neurosci.* 36, 5763–5774. doi: 10.1523/JNEUROSCI.3603-15.2016

Wandell, B. A., Dumoulin, S. O., and Brewer, A. A. (2007). Visual field maps in human cortex. *Neuron* 56, 366–383. doi: 10.1016/j.neuron.2007.10.012

Warburton, M., Mon-Williams, M., Mushtaq, F., and Morehead, J. R. (2022). Measuring motion-to-photon latency for sensorimotor experiments with virtual reality systems. *Behav. Res. Methods*. doi: 10.3758/s13428-022-01983-5. [E-pub Ahead of print].

frontiers | Frontiers in Computer Science

TYPE Hypothesis and Theory
PUBLISHED 11 May 2023
DOI 10.3389/fcomp.2023.1113609

OPEN ACCESS

EDITED BY
James Elder,
York University, Canada

REVIEWED BY
Taiki Fukiage,
NTT Communication Science Laboratories, Japan
Peter König,
Osnabrück University, Germany
Nicholas Baker,
Loyola University Chicago, United States

*CORRESPONDENCE
Christian Jarvers
✉ christian.jarvers@uni-ulm.de

RECEIVED 01 December 2022
ACCEPTED 18 April 2023
PUBLISHED 11 May 2023

CITATION
Jarvers C and Neumann H (2023) Shape-selective processing in deep networks: integrating the evidence on perceptual integration. *Front. Comput. Sci.* 5:1113609. doi: 10.3389/fcomp.2023.1113609

Shape-selective processing in deep networks: integrating the evidence on perceptual integration

Christian Jarvers* and Heiko Neumann

Institute for Neural Information Processing, Faculty for Engineering, Computer Science and Psychology, Ulm University, Ulm, Germany

Understanding how deep neural networks resemble or differ from human vision becomes increasingly important with their widespread use in Computer Vision and as models in Neuroscience. A key aspect of human vision is shape: we decompose the visual world into distinct objects, use cues to infer their 3D geometries, and can group several object parts into a coherent whole. Do deep networks use the shape of objects similarly when they classify images? Research on this question has yielded conflicting results, with some studies showing evidence for shape selectivity in deep networks, while others demonstrated clear deficiencies. We argue that these conflicts arise from differences in experimental methods: whether studies use custom images in which only some features are available, images in which different features compete, image pairs that vary along different feature dimensions, or large sets of images to assess how representations vary overall. Each method offers a different, partial view of shape processing. After comparing their advantages and pitfalls, we propose two hypotheses that can reconcile previous results. Firstly, deep networks are sensitive to local, but not global shape. Secondly, the higher layers of deep networks discard some of the shape information that the lower layers are sensitive to. We test these hypotheses by comparing network representations for natural images and silhouettes in which local or global shape is degraded. The results support both hypotheses, but for different networks. Purely feed-forward convolutional networks are unable to integrate shape globally. In contrast, networks with residual or recurrent connections show a weak selectivity for global shape. This motivates further research into recurrent architectures for perceptual integration.

KEYWORDS

convolutional networks, shape, Gestalt, recurrent connections, deep learning, perceptual grouping

1. Introduction

The success of deep neural networks has led to a new convergence of research in Computer Vision and Neuroscience (Kriegeskorte, 2015). Many motifs in neural network architectures have been loosely inspired by the brain. For example, the local filters used in convolutional neural networks resemble connections in the ventral visual stream of primate cortex. This analogy is fruitful for both sides: on the one hand, further biological inspiration may help improve deep networks by bringing them closer to the robustness and flexibility of biological vision (Medathati et al., 2016). On the other hand, deep networks can serve as models for neuroscience, allowing researchers to implement and test new hypotheses (Kriegeskorte, 2015; Cichy and Kaiser, 2019; Richards et al., 2019). Several studies have

used convolutional neural networks as models of the visual system and been successful at predicting responses (Cichy et al., 2016; Yamins and DiCarlo, 2016; Zhuang et al., 2021) and representational geometries (Khaligh-Razavi and Kriegeskorte, 2014) in the ventral stream, culminating in efforts to find neural network architectures that predict brain data well (Schrimpf et al., 2020).

However, several pieces of evidence suggest that neural networks classify images according to very different criteria than primate vision. Small changes below the human perceptual threshold can turn an image into an adversarial example, which networks classify wrongly with high confidence (Szegedy et al., 2014). More generally, deep networks are much less robust to image corruptions than humans (Geirhos et al., 2018).

Clearly, there are some parallels between deep networks and primate vision, but also crucial differences. The question is: what are the similarities, precisely? And what causes the differences? The answers to these questions are relevant for Neuroscience, since they will circumscribe the extent to which deep networks are useful as models of primate vision. They are also relevant to Computer Vision, since they may help improve neural networks, for example, by making them more robust against adversarial examples and image distortions.

To identify similarities and differences, it is useful to start with key properties of primate vision and test whether deep networks share these properties. One fundamental aspect of human vision is the perception of shape. Distinguishing different objects in our environment, understanding where the boundaries of each object lie, and how they are arranged in our 3D environment are some of the main functions of primate vision. But what about deep networks? A large body of work has been dedicated to this question in recent years - with conflicting results. While neural networks seem to classify images preferentially by shape in some studies (Ritter et al., 2017; Tartaglini et al., 2022), other experiments show that networks are biased toward texture (Baker et al., 2018; Geirhos et al., 2019a). In some papers they can be made sensitive to shape by changes to the training process (Geirhos et al., 2019a; Hermann et al., 2020), whereas in others they are unable to learn about shape (Baker and Elder, 2022). How should these conflicting findings be interpreted?

In this paper, we review research that investigates shape processing in deep networks trained to classify images[1] and compares it to primate vision. Our goal is to reconcile results that appear contradictory. We argue that this is due to differences in the experimental methods, which focus on different aspects of shape processing. Some methods test whether networks are sensitive to the global arrangement of object parts, others also treat local shape cues (e.g., corners) as shape information. Some methods assess whether networks *can* use shape cues, while others test whether networks *prefer* shape over other features.

Taking these distinctions into account, we propose two alternative hypotheses that explain the evidence from previous studies: firstly, networks only use local shape cues, but are not sensitive to global shape. Secondly, networks may process shape in intermediate layers, but discard it in the final decision layers. We argue that a combination of experimental approaches is necessary to test these hypotheses and present evidence from such an experiment, bridging previous studies. According to the results, both hypotheses may be correct, but for different network architectures. While purely feed-forward networks are unable to process global shape, networks with residual or recurrent connections show some selectivity for global shape in intermediate layers, but discard this information at later stages in the network hierarchy. This opens up new opportunities for research on recurrent grouping in deep networks.

2. Do convolutional networks process shape? Conflicting evidence

Human shape perception is a complicated process. According to current theories in neuroscience, object features including cues about *local shape* (such as corners or boundary contours) are initially extracted in a feed-forward pass through the ventral visual stream, establishing a base representation (Roelfsema and Houtkamp, 2011; Elder, 2018). These cues may be sufficient to support object recognition in simple scenarios. For example, to recognize a cat it may be enough to see the distinctive local contours of its ears. The ability to recognize objects quickly in such simple circumstances has been dubbed core object recognition (Afraz et al., 2014; but see also Bracci and Op de Beeck, 2023). However, in more difficult viewing conditions (e.g., partial occlusion and multiple objects), the brain has to group parts of the object together and segment the object from the background. For this kind of robust, flexible processing of object shape, lateral and feedback connections are crucial (Roelfsema and Houtkamp, 2011; Elder, 2018), as they support the grouping of object contours (Grossberg and Mingolla, 1985, 1987; Tschechne and Neumann, 2014), assignment of border ownership (Craft et al., 2007), and segmentation of the object from its background (Self and Roelfsema, 2014). Importantly, this recurrent grouping is highly sensitive to the relative arrangement of object parts, the *global shape*. The set of rules by which object parts are grouped together has been studied extensively in Gestalt psychology and its successors (Wagemans et al., 2012). The cumulative effect of this grouping is that the object is perceived as a unified whole, a Gestalt.

Do deep network represent shape in a similar manner? Initial work tried to address the question directly by comparing responses and representations between deep networks and primate vision. Kubilius et al. (2016) tested whether human participants and deep networks could recognize objects just by their silhouette. Since the silhouette only contains information about object shape, this would indicate shape processing. Indeed, both human participants and deep networks could recognize some object classes by shape and their performance was correlated. Deep networks performed more

1 We focus on networks trained for image classification or object recognition, because (1) most work on shape processing in deep networks has focused on this task and (2) object shape is an important factor in the way humans recognize objects. However, recognizing objects is only one of many capabilities of human vision. It is possible (and highly likely) that the way humans perceive shape is influenced by the many other visual behaviors they exhibit. Looking at shape processing in deep networks trained for other tasks is an interesting direction for future research, but beyond the scope of this paper.

poorly on objects that were hard to classify for human participants. In addition, Kubilius et al. (2016) tested how networks represented images of artificial shapes. Notably, the outputs of hidden layers were more correlated for images of shapes that humans judged to be similar than for shapes that were physically similar. Similarly, Kalfas et al. (2018) used representational similarity analysis (Kriegeskorte et al., 2008; Diedrichsen and Kriegeskorte, 2017, see also Section 2.4) to show that representations in deep networks were highly similar to neural activity in macaque inferotemporal cortex when viewing artificial 2D shapes and to human similarity judgements about the same stimuli.

While these results are encouraging and support the view that deep networks can serve as models of the ventral stream, they do not tell us much about *how* deep networks process shape. For example, the similarity between network representations and human or primate vision may be because the networks extract similar features as the initial feed-forward sweep through the ventral stream. This is supported by the fact that stimuli were presented for only 100 ms by Kubilius et al. (2016), leaving little time for recurrent processing (Thorpe et al., 1996). However, the human similarity judgements reported by Kalfas et al. (2018) were based on unrestricted viewing, so here the match to deep networks might reflect that they are sensitive to global shape.

Since we are far from understanding human vision perfectly, direct comparisons between humans and deep networks cannot answer these detailed questions. Instead, several studies have designed experiments to probe the characteristics of shape processing in deep networks directly. These experiments can be roughly subdivided into four different categories (see Figure 1):

1. Classification of diagnostic stimuli,
2. Classification of cue conflict stimuli,
3. Triplet tests,
4. And representation analysis.

Notably, each category operationalizes the concept "shape" in a different way and tests different aspects of shape processing. This can cause apparent contradictions when comparing results. However, the results within each category are relatively consistent. To demonstrate how the apparent contradictions can be resolved, we look at each experimental approach in turn. We summarize their respective findings and analyze what the advantages and limitations of each approach are.

2.1. Classification of diagnostic stimuli

One way to test how deep networks process shape is to create custom images in which shape information is isolated from other confounding factors, or in which shape information is manipulated selectively (see Figure 2). If a network is able to correctly classify images in which all information except shape is removed (for example silhouettes), then the network must be using features that encode shape. Conversely, if manipulating the shape information (e.g., by shuffling image patches) affects the network output, this indicates that the network used this information to classify the image.

As noted above, Kubilius et al. (2016) showed that convolutional networks can recognize some objects by their silhouette, which indicates that they use at least some shape information. Baker et al. (2018) replicated this result, but also tested a wider range of diagnostic images (as well as cue conflict stimuli—see Section 2.2). The neural networks tested (AlexNet and VGG-19) performed much worse on line drawings of objects, which contain at least as much information about object shape as silhouettes. The only difference is that in a line drawing the interior of an object has the same color as the background, whereas the interior of a silhouette is filled with a uniform color that is different from the background.

In addition, Baker et al. (2018) tested what kind of shape information the networks used to classify silhouettes: local or global shape. Human perception of shape partially uses local shape cues, such as orientation or curvature (Elder, 2018). For example, the characteristic shape of cat ears may be helpful in recognizing a silhouette image as a cat. However, the evidence from these local shape cues is not simply accumulated. Instead, human shape perception is strongly influenced by the global arrangement of these local cues, for example whether the parts of an object are in the correct positions relative to each other and whether they form a closed contour (Wagemans et al., 2012).

In order to test whether neural networks primarily rely on local or global shape information, Baker et al. (2018) modified the silhouette stimuli in two ways. First, they created scrambled silhouettes (see Figure 2E), in which the original silhouette was cut apart and pasted back together in a different arrangement. This largely conserved local shape cues but completely altered the global shape. Second, they manipulated the local boundaries of the original silhouettes by adding a saw-tooth effect (see Figure 2F). This changed the local shape features, but left the global arrangement intact. Human participants showed low accuracy on the scrambled silhouettes and high accuracy on the locally perturbed silhouettes, indicating that they primarily rely on global shape. In contrast, deep networks performed better on the scrambled silhouettes than on the locally perturbed silhouettes, indicating that they relied mainly on local cues and combined them like a bag-of-features model, in line with Brendel and Bethge (2019).

Similarly, Baker and Elder (2022) compared the performance of human participants and deep networks on silhouettes and tested several manipulations that altered the global shape. In fragmented silhouettes (see Figure 2G), the shape was cut in half and the two halves were moved apart. In Frankenstein silhouettes (see Figure 2H), the upper half of the silhouette was flipped horizontally and both halves were pasted back together. Finally, Baker and Elder (2022) also used vertically inverted versions of all these stimuli. The performance of humans and deep networks was worse on fragmented silhouettes, which introduced a new local shape feature (the horizontal cut). However, humans also performed worse on the Frankenstein stimuli, in which global shape was altered while keeping local cues largely identical. Deep networks performed equally well on Frankenstein stimuli as on the original silhouettes, indicating that they did not rely on global shape. This effect

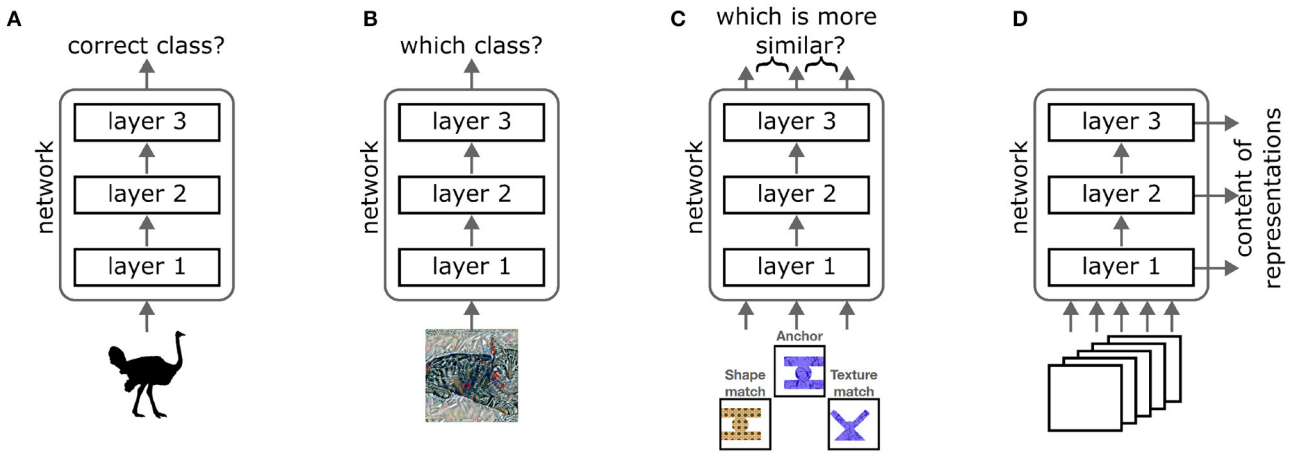


FIGURE 1
The different types of experiments used to assess shape processing in neural networks. **(A)** Diagnostic stimuli restrict the amount of information available in an image. For example, silhouettes only show the shape of an object. If a network can classify such stimuli correctly, it must be able to use the available information. **(B)** Cue conflict stimuli combine features of two different classes, for example, the shape of a cat with the texture of a bicycle. If the network chooses one of these two classes, this indicates that it weights the respective feature more strongly. **(C)** In triplet tests, an anchor stimulus and two matches are presented to the network. The matches differ from the anchor along two different feature dimensions. By testing which representations are more similar, one can assess which features the network uses to group stimuli. **(D)** Methods to analyze network representations record the outputs of intermediate layers across many images and asses which information is present in the representations. Image credit: Silhouette reproduced from Baker et al. (2018), Figure 18 (CC-BY 4.0 attribution license). Cue conflict image reproduced from: https://github.com/rgeirhos/texture-vs-shape (CC-BY 4.0 attribution license). Triplet stimuli reproduced from Tartaglini et al. (2022) (CC-BY 4.0 attribution license).

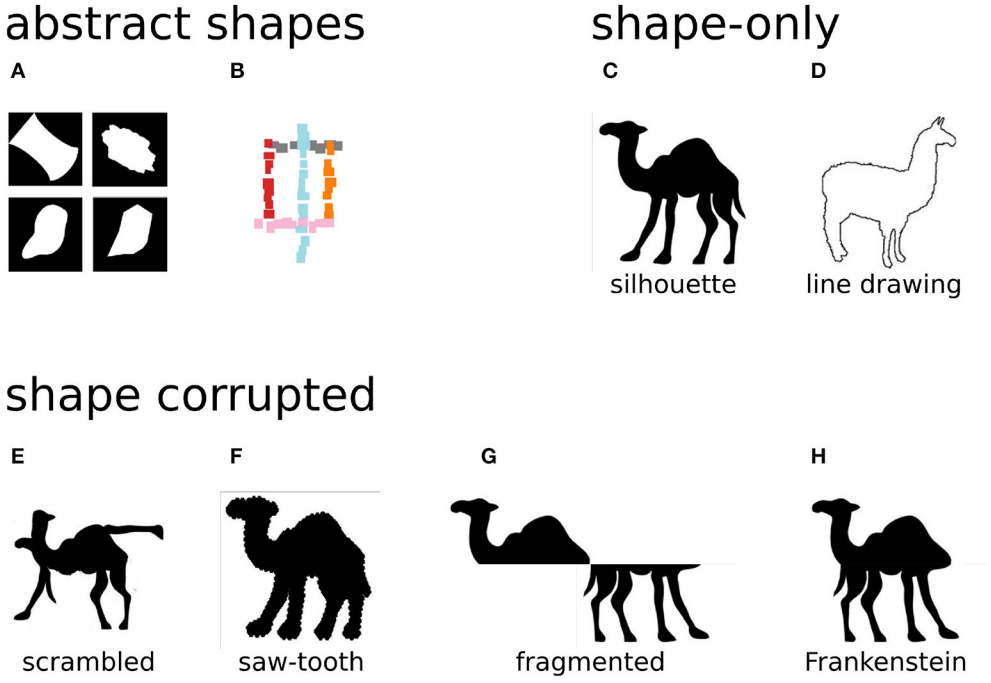


FIGURE 2
Diagnostic stimuli isolate or manipulate shape information. Artificial, abstract shapes allow tight control over which shape information is present: **(A)** Kalfas et al. (2018) generated shapes in four categories: regular (top left), complex (top right), simple curved (bottom left), and simple straight (bottom right). **(B)** Malhotra et al. (2022a) designed artificial shapes that could be classified according to their shape or another feature (e.g., the whether a red segment is present). Shape-only stimuli depict real-world objects but remove all information except the shape, for example, by extracting a silhouette **(C)** or line drawing **(D)**. Shape corrupted stimuli additionally manipulate the shape information, for example by scrambling **(E)** the silhouette or distorting its boundary **(F)**. Similarly to scrambling, Baker and Elder (2022) used fragmented silhouettes **(G)** and re-aligned the parts to create "Frankenstein" stimuli **(H)**. Image credits: **(A)** reproduced from Kalfas et al. (2018) (CC-BY 4.0), **(B)** reproduced from Malhotra et al. (2022a) (CC-BY 4.0), **(C–F)** reproduced from Baker et al. (2018) (CC-BY 4.0), and **(G, H)** adapted from Baker et al. (2018) (CC-BY 4.0).

was constant across different network architectures, including the biologically motivated, recurrent CORnet architecture (Kubilius et al., 2019), a ResNet model trained on stylized ImageNet to be more sensitive to shape (Geirhos et al., 2019a), and vision transformers, which use self-attention to potentially integrate information globally across the image and have been argued to resemble human vision more closely than convolutional networks (Tuli et al., 2021).

Inverting the silhouettes horizontally reduced the performance for humans and deep networks (Baker and Elder, 2022), but in humans this effect was less strong for inverted Frankenstein stimuli, indicating holistic processing. For deep networks, there was again no difference between original and Frankenstein silhouettes.

In sum, these experiments indicate that neural networks trained on ImageNet are not sensitive to global shape of silhouettes. But what can we infer from this about how neural networks process natural images? A potential problem arises due to domain shift. The networks examined in Kubilius et al. (2016), Baker et al. (2018), and Baker and Elder (2022) were trained on natural images, silhouettes were not part of their training set. Deep networks typically transfer badly to data outside of their training distribution. It is conceivable that a network uses both global and local shape in its training domain (natural images), but when faced with a new domain (silhouettes) only some of the features it has learned transfer well enough to enable classification.[2] To rule out the possibility that networks use global shape, one would need to test them on diagnostic stimuli inside their training domain.

Experiments closer to this requirement were conducted by Baker et al. (2020), who used transfer learning to have networks pre-trained on ImageNet classify images of circles and squares. They then tested the network on images of squares composed of small half-circles and circles composed of small corner-like wedges. While this also presents a shift away from the training distribution, it is less drastic than the shift from natrual images to silhouettes. Importantly, a network that is only sensitive to local shape features might distinguish squares from circles based on their sharp corners - and should therefore miss-classify a circle made up of small corner-like elements. This is exactly what Baker et al. (2020) observed with networks trained on simple circles and squares. When they instead trained the networks on circles and squares made up of more diverse local elements (like crosses, tilde signs, or thicker lines), networks responded in line with the global shape of the stimulus: circles made up of small corners were classified as circles, squares made up of half-circles were classified as squares. However, when the networks were tested on shapes made up of small, randomly oriented line segments, performance was largely random (indicating that the networks still relied on local shape) and the networks treated fragmented squares or circles the same as whole shapes (indicating that changes to global shape did not matter). Baker et al. (2018) concluded that the networks still used local cues, but at a slightly larger scale: they ignored the very small elements (corners or half-circles) and instead checked whether the overall orientation was constant (as for squares) or changed gradually (as for circles). They did not use global shape.

Similarly, Malhotra et al. (2020, 2022a) trained networks on custom datasets for which the network could either learn to classify by shape or by another feature. Malhotra et al. (2020) added noise or a single diagnostic pixel to natural images. The statistics of the noise (e.g., the mean) or the color of the single diagnostic pixel indicated the image class, so the network could classify the image either by the appearance of the depicted object, or by the noise. The networks relied heavily on the noise or pixel features, showing drastically reduced or random performance on clean images. Even if the manipulations were restricted to a subset of the training classes, so that the network had to use object appearance to classify the remaining classes correctly, networks relied on the noise/pixel features for as many classes as possible.

Malhotra et al. (2022a) ran similar experiments with completely artificial stimuli and compared the behavior of human participants and deep networks. The stimuli could be classified according to shape or one other feature (e.g., the color of one image patch, see Figure 2B). Participants almost always learned to classify the objects by shape, except for one experiment where the other feature was the color of a large part of the stimulus. When shape was not available as a cue, participants struggled to learn the task at all, even when they were told what the diagnostic feature was. In contrast, neural networks systematically preferred all other features over shape. They appeared to learn some shape information, since their accuracy was above chance when the other feature was removed. However, when faced with a stimulus where shape and the other feature were in conflict, the networks always classified according to the other feature.

Taken together, experiments with diagnostic stimuli show that neural networks are sensitive to some shape information and can use it to classify silhouettes. However, they rely on local shape cues rather than global shape and if they have the choice between shape and another informative feature, they typically use the other feature, such as local color, texture, or even noise statistics.

While this evidence seems compelling, it has to be taken with a grain of salt. Diagnostic stimuli are usually far from the training distribution, so we cannot just assume that neural networks behave identically on the natural images they were trained on. In addition, these experiments rely on the classification output of the networks.[3] It is possible that networks extract shape information in earlier layers, but largely discard it in the final layers because other features are more predictive (see Section 2.3). Conversely, just because networks are able to classify some diagnostic stimuli according to local shape cues, this does not mean that they rely on these cues when classifying natural images. Whether they do can be tested using cue conflict.

2 Hosseini et al. (2018) proposed using negative images (i.e., images with inverted intensity values) to assess shape processing. Negative images may suffer less from domain shift than silhouettes, but have the disadvantage that they do not eliminate texture information, so their diagnostic value is less clear.

3 Baker et al. (2020) also examined correlations among activities in earlier layers. However, these seemed to be dominated by input similarity and did not reveal much about shape processing.

2.2. Classification of cue conflict stimuli

Whereas diagnostic stimuli can be used to assess whether neural networks *are able to* use shape information, cue conflict stimuli are designed to test whether they *do* use this information. In order to test whether a network classifies an image by shape or by another feature, for example texture, one can generate an artificial image with the shape of one class, but the texture of another. This puts the two features or cues in conflict. By observing which class the network predicts, one can assess which feature it relies on more strongly.

Baker et al. (2018) filled silhouettes of one class with surface content of another (see Figure 3A). For example, the outline of a camel was filled with the stripes of a zebra's fur. Deep networks had a low accuracy on this dataset, but still identified the shape or texture of some images correctly. Notably, if the silhouette was from a human-made object, the networks had a higher likelihood of identifying the class of the shape, but if the silhouette was from an animal, the networks were more likely to identify the texture. This may be due to the fact that many human-made artifacts have clear edges and corners, i.e., very distinctive local shape cues, but relatively homogeneous surfaces with less texture information.

Since Baker et al. (2018) created images by hand, they could only test a limited range of conflict stimuli. Geirhos et al. (2019a) used neural style transfer (Gatys et al., 2016) to create stimuli with the shape of one class and the texture of another (see Figure 3B). They tested human participants and convolutional networks (AlexNet, VGG-16, GoogLeNet, and ResNet-50) on 1,280 images, each of which belonged to one of 16 classes. Shapes and textures were counterbalanced in their frequency of presentation. The authors defined a measure of *shape-bias* and *texture-bias* as the fraction of images classified by shape (or texture, respectively) out of the total number of images classified according to either shape or texture. The measure excludes images that were not classified correctly according to either cue. While humans exhibited a strong shape-bias, neural network mostly classified according to texture.

This definition of shape-bias as the fraction of shape decisions made on cue conflict stimuli derived by style transfer has been adopted widely in the deep learning community and has been the main target of attempts to improve the way neural networks process shape. For example, Geirhos et al. (2019a,b) showed that training networks on randomly stylized images, for which the style/texture is no longer predictive of the class, can increase shape bias and that this increased shape bias also leads to higher robustness against image distortions such as noise. While training only on stylized images led to reduced performance on natural images, training on a mix of natural and stylized images led to good performance on both, as well as increased robustness. Hermann et al. (2020) showed that networks could be explicitly trained to use the shape or texture cue and that changes to the training procedure—longer training with stronger augmentations and less aggressive cropping—could lead to higher shape bias. In contrast, changes in architecture (e.g., using an attention layer or the biologically inspired CORnet model) did not have a clear effect. Other methods to improve the shape bias include mixing in edge maps as training stimuli and to steer the stylization of training images (Mummadi et al., 2021), applying separate textures to the foreground object and the background (Lee et al., 2022), penalizing reliance on texture with adversarial learning (Nam et al., 2021), training on a mix of sharp and blurry images (Yoshihara et al., 2021), adding a custom drop-out layer that removes activations in homogeneous areas (Shi et al., 2020), or adding new network branches that receive preprocessed input like edge-maps (Mohla et al., 2022; Ye et al., 2022).

Notably, most of these adjustments have to be carefully tuned, otherwise the networks with improved shape bias perform worse on natural (non-stylized) images. In addition, improvements in shape bias do not always lead to improvements in robustness (Mummadi et al., 2021). We should therefore be cautious in interpreting these results: a higher shape bias may not mean more human-like understanding of shape. As a case in point, Tuli et al. (2021) included shape bias in a larger comparison of convolutional networks and vision transformers (ViT) to human vision. While the ViTs had a higher shape bias, the error pattern (which classes were mistaken for which other classes) of ResNets resembled that of human participants more closely.

Another potential problem comes from the method to create the cue conflict stimuli in most studies. Neural style transfer (Gatys et al., 2016) attempts to preserve the content (i.e., the shape) of one image while applying the texture of another by performing gradient descent with a content loss and a style loss. The *content loss* ensures that the activations of one layer in a deep network are kept close to the activations for the content image. Typically, a layer higher up the network hierarchy is used in order to capture high-level semantic features. The *style loss* is computed across several convolution layers to capture both high- and low-level image features. For each layer, it penalizes the distances between the Gram matrix of activations in that layer for the style image and the image that the style is transferred to. This means that the stylized image will elicit the same correlations between feature detectors in that layer of the network as the style source image. The result is an image in which structures of the content image will still be recognizable to humans.

For example, if the content source shows a house, the outline or shape of the house will be largely intact. However, the color and surface properties will be taken from the style source. For example, if the style source is a painting, the walls may be painted in brush-strokes. At least, this is what the resulting image looks like to a human observer. The key point to keep in mind for this discussion of shape bias is that a stylized image is generated by gradient descent with respect to activations in a neural network. Since neural networks can be sensitive to image features that are not perceptible to humans (Szegedy et al., 2014), this process might introduce features that strongly bias neural network responses, but that are not visible to a human observer. Conversely, it could destroy shape features that networks use to classify images - thereby causing the low shape bias.

In summary, despite these caveats, evidence from cue conflict largely corroborates the findings from diagnostic stimuli: neural networks do not classify images according to object shapes. Rather, they rely on texture cues. However, this preference for texture can be weakened by modifications to the network architecture or training procedure.

In contrast to experiments with diagnostic stimuli, cue conflict tests do not distinguish between local and global shape information.

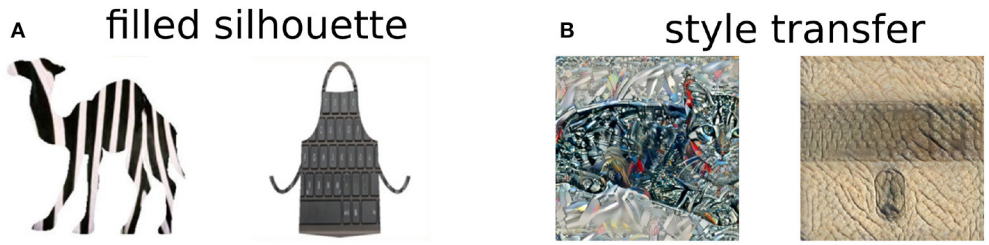


FIGURE 3
Stimuli used in cue conflict experiments. **(A)** Silhouettes filled with surface texture of another class. **(B)** Cue conflict generated using style transfer, using an image of one class as content or shape and an image of another class as style. The left image has shape "cat" and texture "bicycle." The right image has shape "keyboard" and texture "elephant." Image credit: **(A)** reproduced from Baker et al. (2018) (CC-BY 4.0), **(B)** reproduced from: https://github.com/rgeirhos/texture-vs-shape (CC-BY 4.0).

Thus, it is not clear whether improvements in shape bias are due to an increased reliance on local shape cues, or because networks learn to integrate shape information globally [though the results from Baker and Elder (2022) indicate the former, see Section 2.1].

In addition, just like experiments with diagnostic stimuli, cue conflict tests rely on the classification output of the network. This may skew the results. For example, it is conceivable that a network trained on ImageNet learns to use both shape and texture to classify natural images. When faced with a cue conflict, it has to base its decision on one of the two features. It may prioritize texture for various reasons, for example because the object surface takes up a larger part of the image than the object outline such that the texture evidence out-votes the shape evidence. To avoid this potential confound, it is necessary to examine shape bias without relying on the classification output alone. This can be achieved using triplet tests.

2.3. Triplet tests

Shape bias in human participants has been extensively studied in cognitive psychology. For example, when learning new words children tend to group objects by shape, rather than texture or size, exhibiting a shape bias, which increases with age (Landau et al., 1988). In order to control for response bias, Landau et al. (1988) adopted a forced-choice procedure: they showed participants one object (the standard) and then had them choose among two other objects that differed from the standard by different features. For example, one might have a different size than the standard, but the same shape. The other had a different shape, but the same size. The participants had to choose the object that they thought belonged to the same category as the standard.

Ritter et al. (2017) adopted an analogous procedure for testing the shape bias of neural networks. They used a probe image that showed an object, as well as color and shape matches. The color match showed an object of the same color as the probe, but with a different shape. The shape match showed an object with the same shape as the probe, but a different color. The authors computed the cosine distance between the activation of the final layer of an Inception network (before applying the softmax) for the probe image and the activation for each match image. If the representation distance between probe and shape match was smaller than between probe and color match, this was counted as a decision for shape. Notably, for the Inception network and for matching nets (an architecture designed for one-shot classification) the distance between probe and shape match was lowest in most cases: the networks were biased toward shape.

Since this approach uses triplets of images, it is referred to as a triplet task. The term "task" is used in analogy to the forced choice task for human participants, not to classification or other tasks networks are trained for. The networks are not trained for the triplet task. In this sense, the term "triplet test" may be more appropriate.

Feinman and Lake (2018) used this approach to look at the emergence of shape biases during training. They trained small networks on artificial datasets of simple shapes, specifically an MLP with a single hidden layer and a convolutional network with two convolutional layers and one fully connected layer. The authors observed a fast emergence of shape biases. However, since shape was the only feature dimension that was predictive of image classes in their datasets, it is unclear whether the same is true for networks trained on natural images, where color and texture are also predictive of object class.

Since the triplet test is based on similarity of activation patterns, it is not restricted to the output layer of a network. Guest and Love (2019) tested all layers of an Inception network with the same triplets used by Ritter et al. (2017). They observed that lower layers were biased toward color, whereas higher network layers were biased toward shape. They also tested simple artificial stimuli, for which the highest layers were biased toward shape. Notably, the results in the lower layers varied drastically depending on whether stimuli were presented in the same image location or not, indicating that the distance function was dominated by low-level pixel similarity.

These results from triplet tests seem to directly contradict the results from diagnostic stimuli (Section 2.1) and cue conflict (Section 2.2). However, this difference might be due to confounds. For example, the results might be specific to the image triplets tested in Ritter et al. (2017) and Guest and Love (2019). A more direct comparison is enabled by Tartaglini et al. (2022), who performed triplet tests with stylized images like the ones used for cue conflict experiments. Each probe stimulus was a cue conflict image and the texture match was another image with the same texture style, while

the shape match was an image with the same object but a different texture. Interestingly, most networks exhibited a texture bias when tested on standard cue conflict stimuli. However, this changed when the background was masked out. The original shape images showed objects on a white background, but the style transfer procedure also added texture in this background region. Thus, the texture arguably covered a much larger area than the shape object. When the conflicting texture was restricted to the object by masking out the background, all networks exhibited a shape bias. Unfortunately, Tartaglini et al. (2022) did not report the classification-based shape bias measure, so their results cannot be compared to Geirhos et al. (2019a) directly. Nevertheless, their results illustrate that results from the triplet test and cue conflict experiments are generally compatible and that it is important to carefully consider details of the experiment.

Two more important experimental variables that Tartaglini et al. (2022) identified were the spatial alignment and size of the stimulus. Shape bias was generally higher when the object was in the same position in the probe and shape match image. This shows that similarity in the triplet test partially just reflects similarity in pixel space, rather than the processing of features like shape or texture. This point was also raised by Guest and Love (2019) and requires appropriate experimental control. Size also played a role: most networks showed a stronger shape bias for smaller stimuli. This might indicate that networks rely on local shape cues, as indicated by experiments with diagnostic stimuli (Section 2.1). If a network only extracted local shape cues with a certain receptive field size, a smaller object would be covered by this receptive field to a larger degree, increasing the diagnostic value of the features. However, the size effect might also be due to an experimental confound, e.g., because a smaller object means there will be less texturized surface. Notably, even a ResNet with random weights showed a strong shape bias in most experimental conditions (Tartaglini et al., 2022), indicating that the shape bias measured in the triplet test does not necessarily indicate a learned understanding of shape.

In addition, it is unclear whether the shape sensitivity that neural networks show in triplet tests is due to local shape cues or global shape processing. This would require diagnostic stimuli that distinguish between local and global information, but diagnostic stimuli are typically very different from the images a network was trained on (see Section 2.1). Due to this domain shift, it becomes even harder to control for confounds in the similarity-based triplet measure. A step in this direction was made by Malhotra et al. (2022b), who designed triplets of artificial shapes to test if networks represented a relational change, i.e., a change in the relative arrangement of object parts, differently from a coordinate change, which did not change object part relations. Unless explicitly trained to classify a certain type of relational change, networks did not show selectivity for relational changes (i.e., smaller triplet distances). This indicates a lack of global shape processing.

In summary, triplet experiments indicate that deep network encode shape to a higher degree than cue conflict tests reveal. This might mean that networks can use shape information in their decision, but when they are forced to classify a stimulus with conflicting features, they discard shape in favor of another feature. In this view, their capacity for shape processing would be masked by the experimental requirements in classification tasks.

A key advantage of the triplet test is that it can also be applied to earlier layers of the network. Thus, if certain kinds of shape processing were restricted to earlier network layers, this could in principle be revealed by triplet tests. However, since the test relies on direct comparisons between image triplets, it is vulnerable to experimental confounds, such as differences in spatial position, size, etc. This problem can be overcome by methods that analyze the content of representations across larger sets of images.

2.4. Analyzing representations

Two methods that have been used to analyze representations in deep network layers are decoding and representational similarity analysis.

Decoding tests whether a feature is represented in a network layer by training a classifier for that feature. The better the classifier performs, the better the feature must be represented in that network layer. Hermann et al. (2020) trained decoders for the texture and shape classes of cue conflict stimuli for the final pooling layers and fully connected layers of AlexNet and ResNet-50. They observed that both shape and texture could be decoded with high accuracy, indicating that both features were represented. However, while texture was represented equally well across layers, the quality of shape representations decreased across fully-connected layers in AlexNet and after the global average pooling in ResNet-50.

In contrast, Islam et al. (2021) assessed the quality of shape encoding for natural images (not cue conflict stimuli) by decoding segmentation masks for the foreground object. They found that shape could best be decoded from higher convolutional layers, which also contained some information about object class (enabling semantic instead of binary segmentation). However, the authors also noted that the decoder often segmented the shape of an object correctly, but assigned different semantic labels to different object parts, indicating that the global shape of the object was not represented. Islam et al. (2021) also quantified the dimension of shape and texture representations in each layer, i.e., the number of units that were selective to each feature. They assessed this by measuring the mutual information between neuron responses for pairs of cue conflict images that had the same shape or texture, respectively. They noted that in most layers, more neurons were selective for texture than for shape. The dimensionality of shape representations was higher for higher network layers, for deeper networks, and for networks trained on stylized images.

Representational similarity analysis (RSA) captures the overall geometry of representations in a network layer across a range of stimuli. It can also be applied to recordings from biological brains, or to response patterns and even makes it possible to compare different systems (Kriegeskorte et al., 2008; Diedrichsen and Kriegeskorte, 2017). The geometry of representations in a layer is first characterized by recording the distance between each pair of stimuli in a representation dissimilarity matrix (RDM). The geometries of two systems (or of the same system on two sets of stimuli) can then be compared by measuring the distance or correlation between two RDMs.

Kalfas et al. (2018) used this method to compare the representations of artificial two-dimensional shapes (see Figure 2A)

between convolutional networks, recordings from IT cortex of primates, and human similarity judgements. While representations in early networks layers mainly reflected pixel-level image similarity, representational geometries in higher layers were similar to primate and human data. The comparison to pixel-wise similarity is notable, since it overcomes one of the problems of triplet tasks, namely the difficulty of controlling the potential confounding effect of image similarity (Section 2.3). Kalfas et al. (2018) also showed that the similarity to human and primate data did not hold for untrained networks.

Singer et al. (2022) compared representations of photographs, line drawings, and sketches (strongly simplified line drawings, see Figure 4) of the same objects across layers of convolutional networks. Since the contours in the line drawings matched the object edges in the photographs to a high degree, this allowed for a dissociation of shape (which was similar between photographs and line drawings) from surface properties (line drawings and sketches consisted of lines on a white background). In convolutional layers, representations of photographs and drawings were more similar to each other than to representations of sketches. This indicates that representations were more selective to the shape features shared between photographs and drawings than to the surface properties shared between drawings and sketches. In fully connected network layers, this similarity decreased and representations of drawings and sketches were similar instead, indicating selectivity for texture. This decrease in photo-to-drawing similarity was less severe in networks trained on stylized images and could be overcome by fine-tuning to a sketch dataset.

In summary, the evidence from decoding and RSA indicates that networks do encode shape, especially in the higher convolutional layers. This is consistent with observations from triplet tasks. However, in contrast to triplet tasks, which also found high shape bias in fully connected layers, representation similarity analysis and decoding suggest that shape information is discarded in fully connected layers. This could explain why cue conflict experiments consistently find a texture bias.

2.5. Integrating the evidence: the holistic picture

At first glance, the results from different experimental methods seem to contradict each other. For example, cue conflict experiments show that deep networks are biased toward texture, whereas triplet tests indicate that they are biased toward shape. However, these apparent contradictions may be largely due to the fact that each method measures slightly different aspects of shape processing. When these differences are considered carefully, a more complete picture emerges.

Triplet tests, representational similarity analysis, and decoding show that deep networks represent object shape. This is supported by the observation that deep networks can classify silhouettes with some accuracy. However, results from representation similarity analysis and decoding indicate that shape representations are discarded in the last network layers. This is consistent with the tendency of networks to classify cue conflict stimuli by texture, not shape. It is unclear why this reduction in shape information is not evident in triplet tasks, but this might depend on how exactly experimental confounds like pixel-wise similarity are controlled (Guest and Love, 2019; Tartaglini et al., 2022). Thus, one possible interpretation of the data is that deep networks do process object shape, but this information is discarded or down-weighted in the final layers and other features determine the classification output.

While this interpretation resolves most contradictions in the data, it leaves one crucial question open: what kind of shape representations do the intermediate network layers represent? Are they limited to local shape cues, or does a global integration of shape information take place? In experiments with diagnostic stimuli, network responses do not show selectivity for global shape. Adjustments to the training regime may increase the shape-bias measured in cue conflict experiments (Geirhos et al., 2019a; Hermann et al., 2020), but they do not seem to make networks sensitive to global shape (Baker et al., 2020; Baker and Elder, 2022). Thus, a second possible interpretation is that networks are unable to use global shape information. Any shape selectivity shown in triplet tasks, RSA, and decoding is based on local shape cues.

This interpretation may also appear attractive as a source of further analogies to neuroscience. According to current theories of human shape perception, global shape processing relies on recurrence and feedback (Roelfsema and Houtkamp, 2011; Elder, 2018). A lack of global shape processing in feed-forward networks would support this theory. However, some feed-forward network architectures may be able to emulate recurrence (Liao and Poggio, 2016) and some networks incorporate them explicitly (Kubilius et al., 2019). Other network motifs like the global self-attention used in vision transformers (Dosovitskiy et al., 2021) may also enable global grouping of information. Thus, it is important to keep in mind that different networks may process shape differently. So far, most studies on shape processing have focused on one or two network architectures, most commonly AlexNet, VGG, and ResNet. While some studies have explicitly compared different architectures and found that they processed shape similarly (Baker and Elder, 2022), additional systematic comparisons are needed to complete the picture.

The current evidence is insufficient to confirm or falsify either of the interpretations we proposed. To close this gap and to narrow down what type of shape information is represented where and how in which deep network architectures, we think it is necessary to combine the different experimental approaches more explicitly. Each of them offers a unique view of shape processing. To get the full picture, we need to put these views together. For example, diagnostic stimuli offer precise control over the features that are available, while cue conflict or triplet tests allow to asses which of two stimuli a network relies on more strongly. A combined set of stimuli that restricts some cues and puts others in conflict, can give a more nuanced view of which cues a network really uses. Decoding or representational similarity analysis could then be used to track these different cues across layers. No single experiment will characterize shape processing in deep networks and there will not be a single yes or no answer to the question if deep networks classify images according to shape. But by connecting the dots we will be able to understand perceptual organization in deep networks in more detail.

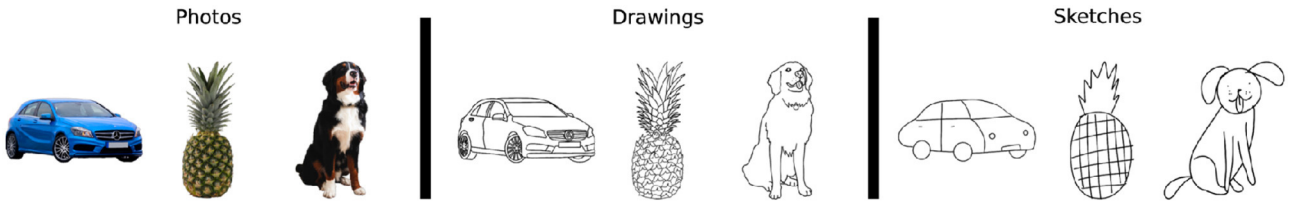


FIGURE 4
Stimuli used in Singer et al. (2022). Line drawings closely match the photographs, such that the represented shape information is similar while the surface properties differ. Line drawings and sketches have similar (absent) surface properties, but show different degrees of shape information. Image credit: reproduced from Singer et al. (2022) (CC-BY 4.0).

3. Experiments—Testing for local and global shape in intermediate representations

We have proposed two hypotheses that can explain previous findings on shape processing in deep networks:

> **Hypothesis 1**: Deep networks trained to classify images are not sensitive to the global arrangement of object parts. Any shape selectivity they exhibit (e.g., in triplet tasks or after shape-biased training) relies on local shape cues, e.g., characteristic parts of the object outline.
> **Hypothesis 2**: Deep networks are sensitive to some shape cues. However, they rely more strongly on other features like texture when classifying images. As these other features are more important or easier to discover in the training set, shape features are down-weighted in the final layers of the network. The networks appear to discard the shape information.

To test these hypotheses, we need to assess to what extent intermediate layers represent shape, and whether the representations reflect local or global shape properties. Baker et al. (2018) and Baker and Elder (2022) designed diagnostic stimuli to dissociate local and global shape processing: silhouettes (which contain only shape information), scrambled or "Frankenstein" silhouettes (in which the global arrangement is disrupted) and jagged silhouettes (in which local shape cues are disrupted). We adopt the same approach. However, instead of manually curating a set of silhouettes, we generate them from natural images that are annotated with segmentation masks. This results in a larger dataset with more variation among object classes and views. In addition, this procedure gives us access to different images of the same object: a natural image, a silhouette, and degraded versions thereof. Due to this one-to-one correspondence, we can compare representations for the different image types using representational similarity analysis, similar to Singer et al. (2022); see also Section 2.4.

Representational similarity analysis compares the similarity structure across items between two representations (Kriegeskorte et al., 2008; Diedrichsen and Kriegeskorte, 2017). We use it to compare representations of diagnostic images with representations of the corresponding natural images in the same network layer. A high representational similarity value (for example, between representations of silhouettes and natural images) means that if the network layer represents two natural images similarly, the representations of the two corresponding silhouettes will be similar as well. This implies that the information available in the silhouette images (i.e., object shape) is relevant for the geometry of the representation in that layer.

This enables us to test our two hypotheses. If hypothesis 1 is true, i.e., networks do not represent global shape, then representations for images that only contain global shape information (i.e., silhouettes in which local shape cues are corrupted) should not be similar to representations of natural images. If hypothesis 2 is true, then there should be significant similarity between representations of shape-only images in early network layers, but not in the final layers of the network. We perform these tests in several networks, to see if differences in architecture affect shape processing.

3.1. Methods

3.1.1. Stimuli

We used images from the PASCAL visual object classes (Everingham et al., 2015). We selected images from the training and validation sets of the 2012 VOC challenge for which detection annotations as well as semantic segmentations were available. We used the detection annotations to remove images with multiple objects and images for which the single object was occluded or truncated. This filtering procedure resulted in 685 images with single, well-visible objects.

To ensure that each object was in the center of the image and all objects were of similar size, we enlarged the bounding boxes provided in the detection annotations by a factor of 1.4 and cropped the image to the resulting window. We resized each image to a resolution of 244-by-244 pixels. Based on these cropped images, we generated a range of diagnostic stimuli (see Figure 5). In foreground images ("fg"), the image background was filled with white color, such that only the object was visible. In silhouette images ("silhouette"), all object pixels were set to black color. To disrupt global object shape, we used a similar method to the "Frankenstein" images in Baker and Elder (2022): we split the image into two halves at the y-coordinate of the center of mass of the silhouette. We flipped the lower half of the image horizontally and re-aligned the edges of the silhouette ("frankenstein"). To disrupt local shape features ("serrated"), we corrupted the silhouette edges,

similar to the jagged silhouettes in Baker et al. (2018). We used a morphological dilation to enlarge the boundary of the object to a width of five pixels. Pixels values in this area were replaced by noise, which we generated by sampling independent, normally distributed values for each pixel, smoothing the result with a Gaussian filter with standard deviation 2, and thresholding at zero.

3.1.2. Networks

We tested a range of networks with different architectural motifs that might influence shape processing. AlexNet (Krizhevsky et al., 2012) and VGG-19 (Simonyan and Zisserman, 2015) are examples of standard convolutional networks, without skip connections or parallel paths. We included GoogLeNet (Szegedy et al., 2015), which uses parallel paths with different kernel sizes, and ResNet-50 (He et al., 2016), which contains residual blocks with skip connections. To test the effect of increasing "shape bias," we also tested a ResNet-50 architecture trained on a mixture of natural and stylized images (Geirhos et al., 2019a). We refer to this network as ShapeResNet. We also evaluated CORnet-S (Kubilius et al., 2019), which has a similar architecture to residual networks but shares weights between residual layers, such that the architecture is equivalent to an unrolled version of a recurrent network (Liao and Poggio, 2016). In addition, CORnet-S was designed to predict activations in the ventral stream of the primate visual system. This makes it an interesting candidate for testing human-like shape perception. We also include BagNet-17 (Brendel and Bethge, 2019), which mirrors the architecture of ResNet-50 but replaces 3×3 convolution kernels in most residual blocks by 1×1 kernels, which restricts the receptive fields of the top-most units to 17×17 pixels. Finally, we evaluate a vision transformer [ViT; Dosovitskiy et al. (2021)], which uses multi-head self-attention between image patches instead of convolutions. As the self-attention operates across the whole image, it could enable the ViT to more efficiently learn global shape properties.

All networks were implemented in PyTorch (version 1.13.0). For AlexNet, VGG-19, GoogLeNet, ResNet-50, and ViT-B-16, we used the implementations and pretrained weights in the torchvision library (version 0.14.0). For BagNet-17 and CORnet-S we used the reference implementations and pretrained weights at: https://github.com/wielandbrendel/bag-of-local-features-models and https://github.com/dicarlolab/CORnet, respectively. For ShapeResNet, we used the weights provided at: https://github.com/rgeirhos/texture-vs-shape. Pretraining for AlexNet, VGG-19, GoogLeNet, ResNet-50, BagNet-17, and CORnet-S was performed on ImageNet-1K with simple image augmentations (random resize and crop, random horizontal flip, and normalization). ShapeResNet was pretrained using the same augmentations, but on a mixture of stylized ImageNet and ImageNet, followed by fine-tuning to ImageNet. ViT-B-16 was also pretrained on ImageNet-1K, but using a more elaborate augmentation pipeline including auto-augmentation, mix-up and cut-mix operations.

3.1.3. Classification

Since all networks we used were trained for ImageNet-1k image classification, their outputs are 1,000-element vectors assigning a probability to each of the 1,000 ImageNet-1k classes. We used the WordNet hierarchy to map each of these outputs to one of the 20 PASCAL VOC classes. Specifically, we translated each PASCAL VOC class to a WordNet synset and collected all ImageNet classes that were descendants of this synset in the WordNet ontology. For example, the ImageNet class "magpie" was mapped to the PASCAL VOC class "bird." For some PASCAL VOC classes, we used hypernyms instead of the original class label in order to capture a wider variety of ImageNet classes (for example, "bovid" instead of "cow"). For each image, we took the top-1 prediction of the network and mapped it onto the respective PASCAL VOC class. If the resulting class matched the label, this was counted as a correct classification. If the prediction was mapped onto the wrong PASCAL VOC class, or if the ImageNet class did not correspond to a PASCAL VOC class (e.g., there is no PASCAL VOC equivalent of the class "envelope"), this was counted as a misclassification. We quantified accuracy as the fraction of correct classifications.

To test if a network was able to use the information in a type of diagnostic image, we compared its accuracy to random performance. Since the number of images per class was not balanced and since some ImageNet classes (which did not have a PASCAL VOC equivalent) were always counted as misclassifications, chance performance depends on the frequency with which the network predicts each class. For example, a network that classifies every image as a random type of bird would have an accuracy of 11.8%, since 81 of the 685 test images were labeled as birds, but a network that classifies every image as a random type of fish would have an accuracy of 0%, since PASCAL VOC does not contain a fish class.

We tested if a network's predictions were significantly more accurate than chance by estimating a null distribution of chance predictions, similar to Singer et al. (2022). For each element in the null distribution, we randomly shuffled the predictions of the network across the 685 images and computed the resulting accuracy. We repeated this procedure 10,000 times. The resulting distribution of accuracies describes how well the network would be expected to perform if it responded randomly, but with the given frequency of each class. To test significance, we computed a p-value as the fraction of elements in the null distribution that were larger or equal to the true (non-shuffled) accuracy of the network. We applied the Benjamini-Hochberg procedure to control the false discovery rate (Benjamini and Hochberg, 1995) for each network. Since this method requires a ranking of p-values, applying it across networks might lead to unwanted interactions (a change in p-value for one network might affect the significance of results for the other networks). We applied a separate FDR-correction to the results of each network and divided the target rate by the number of networks ($0.05/8 = 0.00625$), which corresponds to a Bonferroni-correction across networks.

3.1.4. Representational similarity analysis

We compared representations of different types of diagnostic images using representational similarity analysis (Kriegeskorte et al., 2008; Diedrichsen and Kriegeskorte, 2017).

For each network, we chose several layers of interest. For AlexNet and VGG-19, we looked at each convolutional layer that was followed by max-pooling, as well as the output of the final average pooling and of each fully connected layer. For GoogLeNet,

FIGURE 5
Examples of diagnostic stimuli used in our experiments. Rows show different stimulus types for four example images (one per column). The original image served as a reference for comparisons. Foreground images ("fg") masked out everything except the object of interest. This enables us to estimate how much responses and representations are influenced by the background. In silhouettes, all object pixels were filled with black color, leaving only shape information. Frankenstein stimuli change the global arrangement of object parts, but leave local shape features largely intact. In serrated silhouettes, local shape cues are corrupted, but the global shape remains intact.

we used the outputs of each inception block, average pooling, and the fully connected layer. For ResNets and related architectures, we used the outputs of each residual block, of the average pooling, and the fully connected layer. For ViT, we used the output of each encoder layer, as well as the classification head.

For each layer in a given network, we generated a representational dissimilarity matrix (RDM) for each image type, by calculating the Euclidean distance between the outputs of that layer for each pair of images. We then compared the RDM for each type of diagnostic image (fg, silhouette, frankenstein, and serrated, see Figure 5) to the RDM for the original images. As a measure of similarity, we used Spearman's rank correlation under random tie-breaking (ρ_a).

To estimate the uncertainty of the RSA comparisons, we performed bootstrapping. For each comparison between two RDMs, we performed 1,000 bootstrap runs. In each run, a random subset of RDM indices (i.e., image pairs) was selected and the rank correlation computed over the sub-sampled RDMs. To test whether a given similarity was above chance, we performed a direct bootstrap test, computing the p-value as $(n_{>0} + 1)/N$ where N is the total number of bootstrap runs and $n_{>0}$ is the number of runs with similarity larger than 0. To correct for multiple comparisons, we used the same procedure as for the classification results, controlling the false discovery rate for each network at a level of 0.00625 (Benjamini and Hochberg, 1995). All RDMs and comparisons between

them were computed using the Python package `rsatoolbox` (version 0.0.5).

3.2. Results

3.2.1. Classification of diagnostic images

Classification performance is shown in Figure 6.

All networks perform best on natural images and somewhat worse on foreground images. This indicates that part of their performance relies on features in the background. For example, they may have learned from the dataset that airplanes are often depicted in front of a blue background. All networks perform much worse on shape-only images (silhouettes, frankenstein silhouettes, and serrated silhouettes). Generating silhouettes from masked images removes the texture that defines the region interior of the depicted object. The drop in performance indicates that the networks strongly rely on such information. This is in line with results reported previously about silhouettes and silhouette-derived stimuli that corrupt local or global shape (Baker et al., 2018), though better performance was reported by Baker and Elder (2022). Performance on our stimuli may be lower because the images we used are more challenging since we generated them programatically from a benchmark image set, whereas previous studies curated stimulus sets manually. In addition, our method of computing accuracy (using only the top-1 prediction of the network) is relatively strict.

The networks differ with respect to their performance on the shape-only stimuli. BagNet-17 is the only network that does not classify silhouettes and frankenstein stimuli above chance level. This may either mean that it suffered more strongly from domain shift than other networks, or that it is unable to process shape. It also performs at chance level for serrated stimuli.

AlexNet and VGG-19 perform above chance for silhouettes and frankenstein images, but not for serrated images. This is the pattern of performance expected for networks that use local, but not global shape cues.

GoogLeNet, ResNet-50, Shape-ResNet, CORnet-S, and ViT classify all image types above chance level, indicating that they are able to use shape cues to some extent. To classify frankenstein silhouettes, the networks have to be tolerant to disruptions in global shape, suggesting that they rely on local shape features. Conversely, to classify serrated silhouettes, they have to be robust against disruptions of local shape cues, indicating that they use shape cues at a larger scale than that of the local noise.

3.2.2. Representational similarity analysis

Figure 7 shows the similarities between representations of original images and diagnostic stimuli for each network.

In all networks except for the vision transformer (ViT), the representations of foreground images ("fg") were highly correlated with representations of the original image in all layers. This shows that large parts of the network representations are dedicated to processing the relevant object. Similarities for shape-only stimuli were generally lower. In many layers, representational similarities for diagnostic stimuli are not significantly above zero. This may either mean that shape does not play a big role in these representations, or that these diagnostic stimuli present too much of a domain shift, such that the network cannot interpret them correctly. Both interpretations are consistent with the low accuracy of all networks on diagnostic stimuli.

In ViT, similarities for shape-only stimuli drop to zero in the final layer (the classification head), which matches the hypothesis that shape information is discarded in the final layers. However, the pattern of results in intermediate layers is less clear. Similarities for foreground images decrease throughout the initial encoder layers and are not significantly different from zero in encoder layers 5–9. In encoder layers 5, 6, and 7, none of the similarities are significantly above zero. Similarities for all types of stimuli grow again in the final layers. A possible explanation is that the self-attention mechanism was misled by the large white regions in our images that resulted from masking out the background. Since self-attention aggregates information globally, an image largely devoid of structure may alter the representations in unexpected ways. Note, however, that ViT has the highest accuracy on foreground images out of all networks (65.1%). Thus, the lack of background did not render it unable to make accurate classifications.

For AlexNet and VGG-19, the similarity between original images and shape-only is chance level in most layers. Similarities for silhouettes and frankenstein stimuli are above chance in the final fully connected layers. This is consistent with the results from our classification experiments and with previous studies that found that AlexNet and VGG are not sensitive to global shape (Baker et al., 2018, 2020; Malhotra et al., 2020; Baker and Elder, 2022). However, in layers conv2 to conv4 of AlexNet, similarities for serrated silhouettes are above chance level.

In GoogLeNet, similarities for shape-only stimuli were not significantly above chance in the first two max-pooling layers, which follow after standard convolutions. In all subsequent layers, i.e., inception blocks, average pooling, and the fully connected layer, all similarities were significantly above chance.

ResNet-50, Shape-ResNet, and CORnet-S showed similar patterns of results: in all three networks, similarities for shape-only stimuli were significantly above chance level after the third residual/recurrent block ("layer3" in the ResNets, "V4" in CORnet-S), which is the block with most repeated applications of the residual/recurrent motif. Similarities for some shape-only stimuli dropped back to chance level in the final block ("layer4"/"IT") and the subsequent average pooling layer (original and frankenstein silhouettes for ResNet-50, serrated silhouettes for CORnet-S, and all three types for Shape-ResNet). However, all similarities are significantly above chance in the final fully-connected layers.

In BagNet-17, similarities for silhouettes and frankenstein stimuli were significant in the first and fourth residual block and in the average pooling layer. In addition, the similarity for frankenstein images was also significant in layer 3. However, similarities for serrated silhouettes were only above chance in the average pooling layer. This suggests that the residual layers in BagNet-17 did not extract global shape information, in contrast to the other ResNet-like architectures.

If this is true, how can the significant similarity in the average pooling layer of BagNet-17 be explained? Average pooling discards information about where in the image a certain feature occurred,

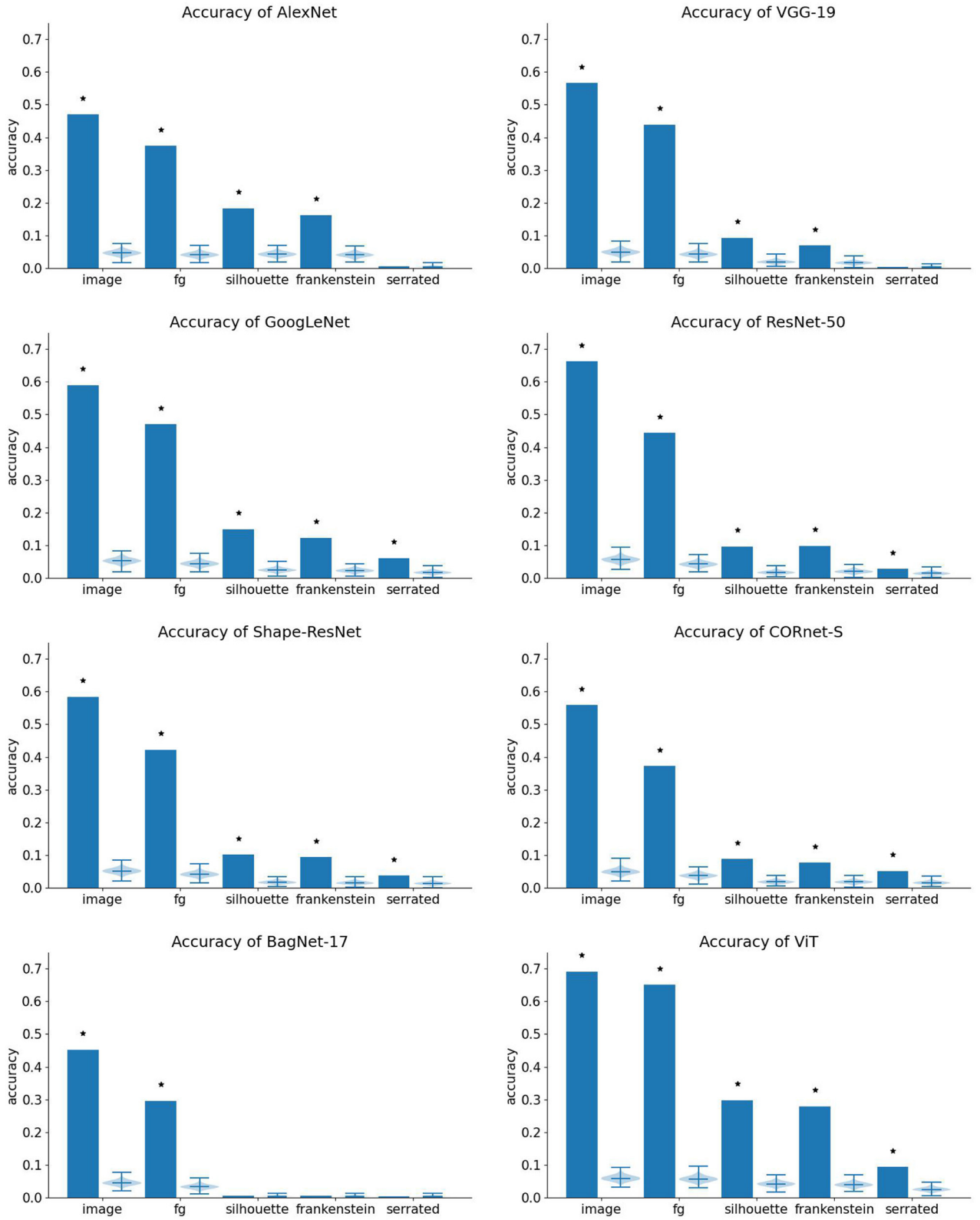


FIGURE 6
Classification accuracy of each network. Bar height indicates performance on the respective dataset. The violin plot to the right of each bar shows the corresponding null distribution. Stars indicate that accuracy is significantly higher than chance.

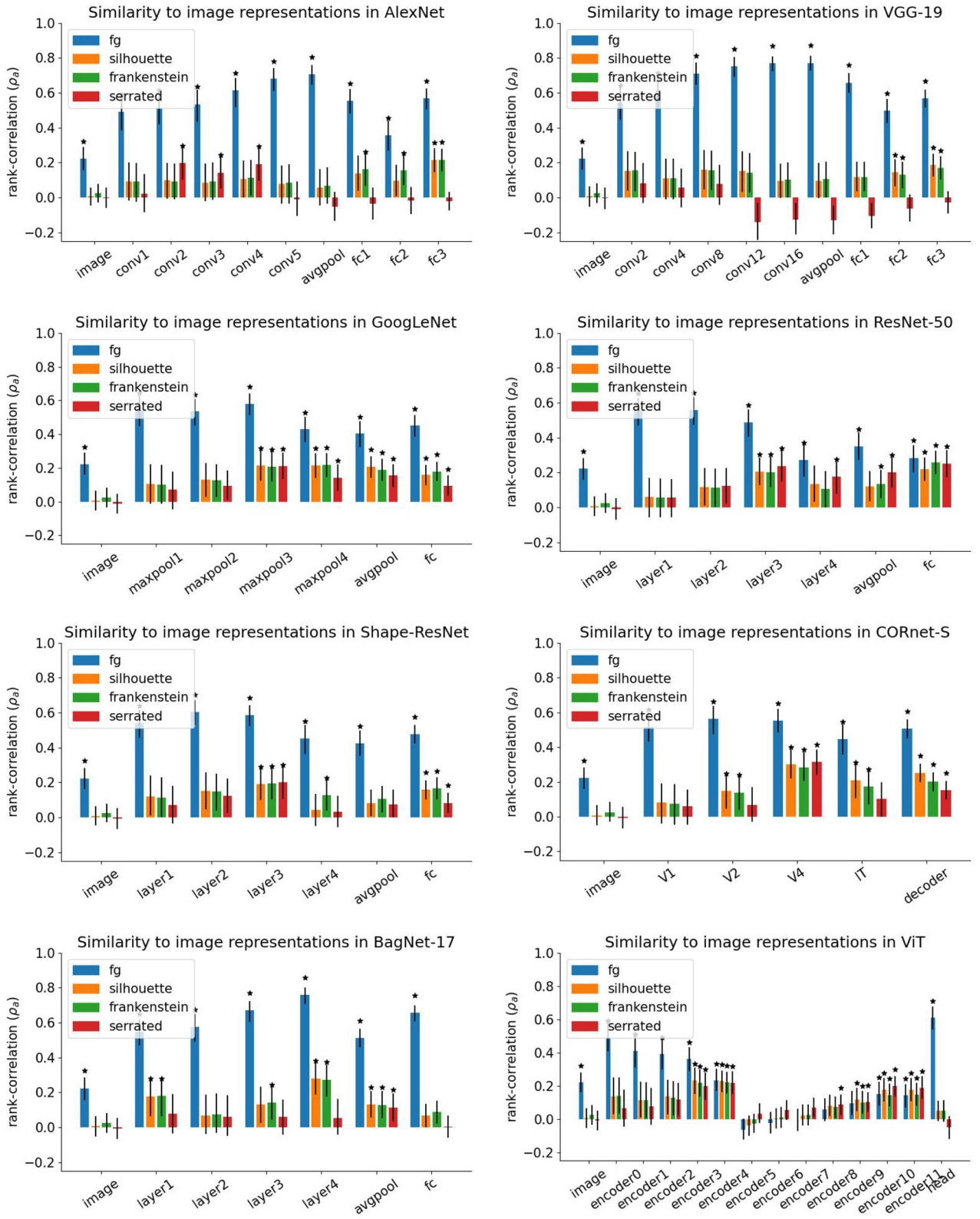


FIGURE 7
Results of representational similarity analysis. Height of bars indicates average rank-correlation over bootstrap runs. Error bars indicate 95% confidence intervals from bootstrap runs. Stars mark similarities that are significantly larger than 0.

since it averages activations of each feature map across all image locations. Therefore, average pooling should make a representation less informative about global shape. At the same time, it may mask a lack of selectivity for global shape in the RSA: if the same feature is detected in two different locations, comparing the resulting representations before pooling would lead to a low similarity, but after average pooling the difference in location vanishes, increasing the similarity. For this reason, the RSA results for average pooling layers should be taken with a grain of salt.

In summary, our results show a difference between network architectures. Classical convolutional networks such as AlexNet and VGG-19 have low shape selectivity in most layers and show a dissociation between serrated silhouettes and other shape-only stimuli without local shape corruptions (original and frankenstein silhouettes). In contrast, GoogLeNet and ResNet-like architectures showed no systematic differences between the shape-only stimulus types, with significant representational similarity for all three types in intermediate inception/residual blocks and in the final layers, though for the ResNet architectures there was a drop in shape selectivity for the final residual blocks. An exception to this pattern was BagNet-17, which had some selectivity for original and frankenstein silhouettes, but not serrated silhouettes, in the residual blocks, but showed no shape selectivity in the fully-connected layer. Finally, the shape selectivity in ViT varied strongly across layers and suddenly dropped in the classification head.

4. Discussion

We have reviewed previous research on shape representations in deep networks and argued that apparent contradictions in their results are largely due to differences in methods. Each method operationalizes the concept "shape" differently and tests different aspects of network processing. Experiments with *diagnostic stimuli* can show that a network is in principle able to use a specific type of shape cue. They allow for fine-grained control over different types of shape cues, for example the availability of local or global shape. However, they induce a strong domain shift, which makes results harder to interpret, and they are limited to the network output. *Cue conflict* experiments can directly compare the influence of two different features over network outputs. But like diagnostic stimuli, they require custom images which leads to domain shift. In contrast, *triplet tests* can be done with natural images and can also directly compare two different features. They can also be applied to intermediate network layers, though this requires strict experimental control of confounding variables. Finally, *representational similarity analysis* and *decoding* offer the most detailed view of representations in intermediate network layers. They can be done using natural images, and typically involve large stimulus sets, making them less vulnerable to domain shift and confounds.

Notably, the strengths of the different approaches are complementary. By combining them explicitly, future research can gain a more detailed understanding of shape processing in deep networks. Some steps in this direction have already been made. Baker et al. (2020) used diagnostic stimuli and also reported correlations between stimulus pairs in intermediate layers, similar to a triplet test. Singer et al. (2022) used RSA to compare representations for photographs, line drawings, and sketches. The latter two stimulus types isolate shape information, similar to diagnostic stimuli. Tartaglini et al. (2022) performed triplet tests with cue conflict stimuli. Nevertheless, many more informative experiments are possible in this combined experimental space.

As an example, we reported results from an experiment that combined representational similarity analysis with diagnostic stimuli designed to distinguish between local and global shape processing. The goal was to test whether (1) intermediate network layers represent global shape features and whether (2) shape features are discarded in final network layers. Both of these hypotheses may explain some apparent contradictions in previous results.

Our results support both hypotheses to some degree. Hypothesis 2 (that networks down-weight shape information in later layers) predicts that a network should have significant representational similarity between original images and shape-only stimuli in intermediate layers, which drops back to chance level in later layers. This is the case for ViT and BagNet, for both of which similarities for shape-only stimuli are at chance in the final fully-connected layer. The results for the ViT should be taken with a grain of salt, however, since it classified shape-only stimuli with above-chance accuracy and the RSA for intermediate layers did not fit either of our hypotheses.

The results for ResNet-50, Shape-ResNet, and CORnet-S also partially match hypothesis 2, as their third residual block showed significant selectivity for shape, which dropped to chance for some types of shape-only stimuli in the final residual block and the average pooling layer. This matches observations by Hermann et al. (2020) that shape could be decoded less accurately after the average pooling layer of ResNet-50. On the other hand, they found the same effect in the fully connected layers of AlexNet, which does not show a similar effect in our RSA.

Hypothesis 1 (that networks only use local shape cues) predicts that networks classify original and frankenstein silhouettes (in which local shape remains intact) above chance level but should fail for serrated silhouettes (in which local shape information is corrupted). AlexNet and VGG-19 match this prediction, both w.r.t. classification and representational similarity in their fully connected layers. This is in line with several previous experiments that used these networks and found a lack of global shape selectivity (Baker et al., 2018, 2020; Baker and Elder, 2022; Malhotra et al., 2022a). However, AlexNet showed above-chance representational similarity for serrated stimuli in early layers. In the top-most convolutional layer and the fully connected layers, this similarity drops back to chance, which either means that AlexNet discards this shape information (as predicted by hypothesis 2) or that other confounding factors play a role, as we discuss below.

In contrast, GoogLeNet, ResNet-50, Shape-ResNet, and CORnet-S classified all stimulus types above chance level and showed significant representational similarity for all shape-only stimuli in intermediate layers. Thus, they represent some shape information, in line with previous results (Hermann et al., 2020; Islam et al., 2021), but which kind of shape information they rely on remains unclear. Since these networks are not affected by the frankenstein manipulation, they seem to be insensitive to global

shape. This confirms the results of Baker and Elder (2022) and is in line with Islam et al. (2021), who observed that decoding with a semantic segmentation objective suffered from errors where different classes were assigned to parts of the same object. This may reflect a lack of global shape understanding.

These four networks were also robust against distortions of local shape cues in serrated silhouettes. Does this mean that they perform some non-local integration of shape features? Alternatively, they may still rely on local cues, but at a larger spatial scale than the corruptions in the serrated images (see e.g., Baker et al., 2020). Since these networks are deeper than AlexNet and VGG-19, their hierarchically organized convolutional layers aggregate input over a larger spatial extent, such that the local noise in the serrated images has a smaller impact. According to this interpretation, our stimulus design would simply not distinguish between local and global shape processing as well as intended. However, this explanation is not entirely convincing: CORnet-S, which exhibited the same shape selectivity as the other ResNet architectures, is considerably less deep and its convolutions have a smaller spatial span than those in VGG-19.

One feature that GoogLeNet and ResNet-like architectures share is the presence of parallel paths. GoogLeNet uses inception blocks, in which the same input is processed by convolutions with different kernel sizes, and the result is concatenated. ResNets and CORnet-S use residual blocks, in which the input to a set of convolutions is added to its output via a skip-connection. As noted by Liao and Poggio (2016), this is equivalent to a one-step temporal unrolling of a recurrent network, in which the convolution spreads information to neighboring locations. Both of these motifs enable a comparison of the image content at one location with the surrounding area. Therefore, they might implement a simple form of lateral grouping, unrolled for a fixed number of steps. This interpretation is supported by three observations. First, the inception modules in GoogLeNet exhibit shape selectivity, but the preceding convolution layers do not. Second, the layers with the clearest selectivity for shape were layer 3 in ResNet-50 and Shape-ResNet and V4 in CORnet-S. These are the blocks which contain the most repetitions of the residual/recurrent motif. Third, BagNet-17 has the same depth as ResNet-50, but replaces the 3x3 convolutions in most residual blocks by 1x1 convolutions, thus restricting the range of lateral connectivity. In contrast to the other ResNets, none of the residual blocks in BagNet-17 had significant representational similarity between original images and serrated silhouettes, suggesting that restricting lateral connectivity impacts non-local shape processing.

If ResNets and GoogLeNet do indeed perform a rudimentary form of lateral grouping, this would constitute another parallel to primate vision, where recurrence is critical for global integration of shape (Roelfsema and Houtkamp, 2011; Self and Roelfsema, 2014; Elder, 2018). The utility of recurrent connections for deep networks has been proposed repeatedly (Kriegeskorte, 2015; Peters and Kriegeskorte, 2021) and several recent network architectures have incorporated it with promising results (Linsley et al., 2018, 2020; Kubilius et al., 2019). Our results suggest that this line of work may enable networks to form more global representations of object shape, reducing the gap between human and machine vision.

Another interesting question for future work is the role of the training objective in shaping the shape selectivity of such networks. Most work to date has focused on characterizing shape processing in deep networks trained to classify images of objects. This is a reasonable starting point, firstly because image recognition on large datasets has been one of the main drivers in the development of deep networks, and secondly because shape is a key factor in how humans recognize objects. When networks learn to classify objects according to human labels, it is tempting to assume that they use the same criteria as humans. The evidence reviewed above clearly shows that this is not the case for shape. Deep networks are still far from using shape information in a human-like manner to recognize objects. A key reason may be that human vision is not limited to object recognition. It supports many other behaviors like visual search, navigation, etc., many of which involve and constrain visual representations of object shape (Ayzenberg and Behrmann, 2022; Bracci and Op de Beeck, 2023). The task of image classification may simply be too under-constrained, allowing deep networks to learn shortcuts (Geirhos et al., 2020). Accordingly, networks trained with self-supervised methods show higher shape bias in some experiments (Hermann et al., 2020; Tartaglini et al., 2022). Future studies examining a broader range of tasks and other types of visual input (e.g., stereo images or video) could deepen our understanding of the constraints that shape the processing of visual shape in hierarchically organized deep networks.

5. Conclusion

Previous research on shape processing in deep networks has yielded conflicting results with some studies showing evidence for shape selectivity, while others showed clear deficiencies. After reviewing the experimental approaches used in these studies, we proposed two hypotheses that can reconcile these results. Firstly, deep networks may rely on local, but not global shape cues to classify objects. Secondly, networks may discard shape information in their final layers and weigh other features more strongly in their classification output, masking their shape selectivity. We tested these hypotheses by combining two of the previously established methods: diagnostic stimuli that restrict the information available in an image, and representational similarity analysis that assesses whether different stimulus sets are represented similarly in a network layer. Our results support both hypotheses—but for different networks. Purely feed-forward convolutional networks like AlexNet and VGG represented local but not global shape. In contrast, networks with inception modules or residual blocks show some selectivity for shape in the presence of local corruptions, which may reflect a simple form of non-local shape processing. This highlights the importance of exploring the effects of different architectural motifs on shape processing. Incorporating more extensive lateral and recurrent connectivity may enable networks to perform iterative grouping and process shape in a more holistic, human-like manner.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: http://host.robots.ox.ac.uk/pascal/VOC/. The code

used for analysis and plotting is available on GitHub: https://github.com/cJarvers/shapebias. The exact version of the code used, and the result files generated, are archived at zenodo: https://doi.org/10.5281/zenodo.7863152.

Author contributions

CJ and HN: conceptualization and writing—review and editing. CJ: investigation, methodology, software, data analysis, visualization, and writing—original draft. HN: supervision. All authors contributed to the article and approved the submitted version.

Acknowledgments

We thank the three reviewers as well as the editor for their kind and insightful feedback. We also want to thank Daniel Schmid, David Adrian, and Irina Jarvers for helpful discussions. The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Afraz, A., Yamins, D. L. K., and DiCarlo, J. J. (2014). Neural mechanisms underlying visual object recognition. *Cold Spring Harbor Symposia Quant. Biol.* 79, 99–107. doi: 10.1101/sqb.2014.79.024729

Ayzenberg, V., and Behrmann, M. (2022). Does the brain's ventral visual pathway compute object shape? *Trends Cogn. Sci.* 26, 1119–1132. doi: 10.1016/j.tics.2022.09.019

Baker, N., and Elder, J. H. (2022). Deep learning models fail to capture the configural nature of human shape perception. *iScience* 25, 104913. doi: 10.1016/j.isci.2022.104913

Baker, N., Lu, H., Erlikhman, G., and Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Comput. Biol.* 14, e1006613. doi: 10.1371/journal.pcbi.1006613

Baker, N., Lu, H., Erlikhman, G., and Kellman, P. J. (2020). Local features and global shape information in object classification by deep convolutional neural networks. *Vis. Res.* 172, 46–61. doi: 10.1016/j.visres.2020.04.003

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Stat. Soc.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bracci, S., and Op de Beeck, H. P. (2023). Understanding human object vision: A picture is worth a thousand representations. *Ann. Rev. Psychol.* 74, 113–135. doi: 10.1146/annurev-psych-032720-041031

Brendel, W., and Bethge, M. (2019). "Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet," in *International Conference on Learning Representations* (New Orleans, LA).

Cichy, R. M., and Kaiser, D. (2019). Deep neural networks as scientific models. *Trends Cogn. Sci.* 23, 305–317. doi: 10.1016/j.tics.2019.01.009

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* 6, 27755. doi: 10.1038/srep27755

Craft, E., Schütze, H., Niebur, E., and von der Heydt, R. (2007). A neural model of figure–ground organization. *J. Neurophysiol.* 97, 4310–4326. doi: 10.1152/jn.00203.2007

Diedrichsen, J., and Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Comput. Biol.* 13, e1005508. doi: 10.1371/journal.pcbi.1005508

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations* (Vienna).

Elder, J. H. (2018). Shape from contour: Computation and representation. *Ann. Rev. Vis. Sci.* 4, 423–450. doi: 10.1146/annurev-vision-091517-034110

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* 111, 98–136. doi: 10.1007/s11263-014-0733-5

Feinman, R., and Lake, B. M. (2018). "Learning inductive biases with simple neural networks," in *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (London: The Cognitive Science Society),1657–1662.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 2414–2423.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., et al. (2020). Shortcut learning in deep neural networks. *Nat. Machine Intell.* 2, 665–673. doi: 10.1038/s42256-020-00257-z

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2019a). "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in *International Conference on Learning Representations*, New Orleans, LA.

Geirhos, R., Rubisch, P., Rauber, J., Temme, C. R. M., Michaelis, C., Brendel, W., et al. (2019b). Inducing a human-like shape bias leads to emergent human-level distortion robustness in CNNs. *J. Vis.* 19, 209c. doi: 10.1167/19.10.209c

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. (2018). "Generalisation in humans and deep neural networks," in *Advances in Neural Information Processing Systems, volume 31*. Dutchess County, NY: Curran Associates, Inc.

Grossberg, S., and Mingolla, E. (1985). Neural dynamics of form perception: Boundary completion, illusory figures, and neon color spreading. *Psychol. Rev.* 92, 173–211. doi: 10.1037/0033-295X.92.2.173

Grossberg, S., and Mingolla, E. (1987). Neural dynamics of surface perception: Boundary webs, illuminants, and shape-from-shading. *Comput. Vis. Graph. Image Proces.* 37, 116–165. doi: 10.1016/S0734-189X(87)80015-4

Guest, O., and Love, B. C. (2019). Levels of representation in a deep learning model of categorization. *biorxiv [Preprint]*. doi: 10.1101/626374

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 770–778.

Hermann, K., Chen, T., and Kornblith, S. (2020). The origins and prevalence of texture bias in convolutional neural networks. *Adv. Neural Inform. Process. Syst.* 33, 19000–19015. doi: 10.48550/arXiv.1911.09071

Hosseini, H., Xiao, B., Jaiswal, M., and Poovendran, R. (2018). "Assessing shape bias property of convolutional neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Salt Lake City, UT: IEEE, 1923–1931.

Islam, M. A., Kowal, M., Esser, P., Jia, S., Ommer, B., Derpanis, K., et al. (2021). Shape or texture: Understanding discriminative features in CNNs. in *International Conference on Learning Representations* (Vienna).

Kalfas, I., Vinken, K., and Vogels, R. (2018). Representations of regular and irregular shapes by deep Convolutional Neural Networks, monkey inferotemporal neurons and human judgments. *PLoS Comput. Biol.* 14, e1006557. doi: 10.1371/journal.pcbi.1006557

Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* 10, e1003915. doi: 10.1371/journal.pcbi.1003915

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Ann. Rev. Vis. Sci.* 1, 417–446. doi: 10.1146/annurev-vision-082114-035447

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 4. doi: 10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Dutchess County, NY: Curran Associates, Inc.),1097–1105.

Kubilius, J., Bracci, S., and Beeck, H. P. O. d. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput. Biol.* 12, e1004896. doi: 10.1371/journal.pcbi.1004896

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., et al. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. *Adv. Neural Inform. Process. Syst.* 32, 12805–12816. doi: 10.48550/arXiv.1909.06161

Landau, B., Smith, L. B., and Jones, S. S. (1988). The importance of shape in early lexical learning. *Cogn. Dev.* 3, 299–321. doi: 10.1016/0885-2014(88)90014-7

Lee, S., Hwang, I., Kang, G.-C., and Zhang, B.-T. (2022). "Improving robustness to texture bias via shape-focused augmentation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (New Orleans, LA), 4322–4330.

Liao, Q., and Poggio, T. (2016). Bridging the gaps between residual learning, recurrent neural networks and visual cortex. *arXiv:1604.03640*. doi: 10.48550/arXiv.1604.03640

Linsley, D., Karkada Ashok, A., Govindarajan, L. N., Liu, R., and Serre, T. (2020). "Stable and expressive recurrent vision models," in *Advances in Neural Information Processing Systems, Volume 33*, eds H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Dutchess County, NY: Curran Associates, Inc.), 10456–10467.

Linsley, D., Kim, J., Veerabadran, V., Windolf, C., and Serre, T. (2018). Learning long-range spatial dependencies with horizontal gated recurrent units. in *Advances in Neural Information Processing Systems 31*, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Dutchess County, NY: Curran Associates, Inc.,152–164.

Malhotra, G., Dujmović, M., and Bowers, J. S. (2022a). Feature blindness: A challenge for understanding and modelling visual object recognition. *PLoS Comput. Biol.* 18, e1009572.

Malhotra, G., Dujmović, M., Hummel, J., and Bowers, J. S. (2022b). Human shape representations are not an emergent property of learning to classify objects. *bioRxiv Preprint*. doi: 10.1101/2021.12.14.472546

Malhotra, G., Evans, B. D., and Bowers, J. S. (2020). Hiding a plane with a pixel: examining shape-bias in CNNs and the benefit of building in biological constraints. *Vis. Res.* 174, 57–68. doi: 10.1016/j.visres.2020.04.013

Medathati, N. V. K., Neumann, H., Masson, G. S., and Kornprobst, P. (2016). Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision. *Comput. Vis. Image Underst.* 150, 1–30. doi: 10.1016/j.cviu.2016.04.009

Mohla, S., Nasery, A., and Banerjee, B. (2022). "Teaching CNNs to mimic human visual cognitive process and regularise texture-shape bias," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Singapore), 1805–1809.

Mummadi, C. K., Subramaniam, R., Hutmacher, R., Vitay, J., Fischer, V., and Metzen, J. H. (2021). "Does enhanced shape bias improve neural network robustness to common corruptions?" in *International Conference on Learning Representations* (Vienna).

Nam, H., Lee, H., Park, J., Yoon, W., and Yoo, D. (2021). Reducing domain gap by reducing style bias. in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Virtual), 8686–8695.

Peters, B., and Kriegeskorte, N. (2021). Capturing the objects of vision with neural networks. *Nat. Hum. Behav.* 5, 1127–1144. doi: 10.1038/s41562-021-01194-6

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., et al. (2019). A deep learning framework for neuroscience. *Nat. Neurosci.* 22, 1761–1770. doi: 10.1038/s41593-019-0520-2

Ritter, S., Barrett, D. G., Santoro, A., and Botvinick, M. M. (2017). "Cognitive psychology for deep neural networks: a shape bias case study," in *Proceedings of the 34th International Conference on Machine Learning—Volume 70, ICML'17* (Sydney, NSW: JMLR.org.), 2940–2949.

Roelfsema, P. R., and Houtkamp, R. (2011). Incremental grouping of image elements in vision. *Attent. Percept. Psychophys.* 73, 2542–2572. doi: 10.3758/s13414-011-0200-0

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., and DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* 108, 413–423. doi: 10.1016/j.neuron.2020.07.040

Self, M. W., and Roelfsema, P. R. (2014). "The neural mechanisms of figure-ground segregation," in *The Oxford Handbook of Perceptual Organization*, Oxford Library of Psychology, ed J. Wagemans (Oxford: Oxford University Press),321–341.

Shi, B., Zhang, D., Dai, Q., Zhu, Z., Mu, Y., and Wang, J. (2020). "Informative dropout for robust representation learning: A shape-bias perspective," in *Proceedings of the 37th International Conference on Machine Learning* (Virtual), 8828–8839.

Simonyan, K., and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *(arXiv:1409.1556). arXiv preprint*. doi: 10.48550/arXiv.1409.1556

Singer, J. J. D., Seeliger, K., Kietzmann, T. C., and Hebart, M. N. (2022). From photos to sketches—How humans and deep neural networks process objects across different levels of visual abstraction. *J. Vis.* 22, 4. doi: 10.1167/jov.22.2.4

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Boston, MA), 1–9.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2014). "Intriguing properties of neural networks," in *International Conference on Learning Representations*. Banff, AB.

Tartaglini, A. R., Vong, W. K., and Lake, B. (2022). A developmentally-inspired examination of shape versus texture bias in machines. *Proc. Ann. Meet. Cogn. Sci. Soc.* 44, 1284–1290. doi: 10.48550/arXiv.2202.08340

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381, 520–522.

Tschechne, S., and Neumann, H. (2014). Hierarchical representation of shapes in visual cortex—from localized features to figural shape segregation. *Front. Comput. Neurosci.* 93. doi: 10.3389/fncom.2014.00093

Tuli, S., Dasgupta, I., Grant, E., and Griffiths, T. L. (2021). "Are convolutional neural networks or transformers more like human vision?" in *43rd Annual Meeting of the Cognitive Science Society: Comparative Cognition: Animal Minds* (London: The Cognitive Science Society), 1844–1850.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., et al. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychol. Bullet.* 138, 1172–1217. doi: 10.1037/a0029333

Yamins, D. L. K., and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365. doi: 10.1038/nn.4244

Ye, Z., Gao, Z., Cui, X., Wang, Y., and Shan, N. (2022). DuFeNet: Improve the accuracy and increase shape bias of neural network models. *Sign. Image Video Process.* 16, 1153–1160. doi: 10.1007/s11760-021-02065-3

Yoshihara, S., Fukiage, T., and Nishida, S. (2021). Towards acquisition of shape bias: Training convolutional neural networks with blurred images. *J. Vis.* 21, 2275. doi: 10.1167/jov.21.9.2275

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., et al. (2021). Unsupervised neural network models of the ventral visual stream. *Proc. Natl. Acad. Sci. U. S. A.* 118, 2014196. doi: 10.1073/pnas.2014196118

frontiers | Frontiers in Computer Science

TYPE Review
PUBLISHED 19 June 2023
DOI 10.3389/fcomp.2023.1124230

OPEN ACCESS

EDITED BY
Mary Peterson,
University of Arizona, United States

REVIEWED BY
Taiki Fukiage,
NTT Communication Science Laboratories, Japan
Cathleen Moore,
The University of Iowa, United States

*CORRESPONDENCE
Joseph S. Lappin
joe.lappin@vanderbilt.edu

RECEIVED 14 December 2022
ACCEPTED 30 May 2023
PUBLISHED 19 June 2023

CITATION
Lappin JS and Bell HH (2023) The coherent organization of dynamic visual images. *Front. Comput. Sci.* 5:1124230. doi: 10.3389/fcomp.2023.1124230

The coherent organization of dynamic visual images

Joseph S. Lappin[1]* and Herbert H. Bell[2]

[1]Department of Psychology, Vanderbilt University, Nashville, TN, United States, [2]Independent Consultant, Mount Dora, FL, United States

Biological vision relies on the intrinsic spatiotemporal structure of a continuously flowing image stream. We review converging psychophysical and physiological evidence about the structure and precision of the perceived spatiotemporal organization of dynamic images. Visual acuity, temporal resolution, and contrast sensitivity have been found to involve (a) motion-produced increases in image contrast, (b) coherent phase relations among temporally varying retinal signals, and (c) physiological preservation of spatiotemporal structure from retina to cortex. Moreover, psychophysical theory and evidence show that the spatiotemporal structure of dynamic retinal images carries precise information for perceiving surfaces and motions—consistent with the corresponding differential structures of spatiotemporal images and environmental surfaces.

KEYWORDS

dynamic images, coherence, perceptual organization, spatial resolution, psychophysics, hyperacuity, contrast, relative motion

Introduction

Vision is a system for acquiring and transmitting dynamic optical information. Even when we fixate on a stationary object, our eyes are constantly in motion. As a result, the location of the image on our retinal photoreceptors is continually changing. Intuitively, these constant changes might seem a threat to vision, analogous to noise. The image changes are not independent, however, and their covariation is a basis for perceptual organization and constitutes information about environmental surfaces, objects, and motions.

Our aim in this article is to show that spatiotemporal structure is essential to the visual organization of retinal images. We briefly summarize psychophysical evidence about visual sensitivity to spatiotemporal variations and describe information provided by those variations. We also discuss mechanisms that may underly the acquisition of this information.

Image motions provide spatial information

Many traditional ideas about vision reflect film-based systems, where image motion causes blur and reduces contrast. Light-induced changes in the photosensitive molecules in film are mainly integrative and independent among neighboring elements. In the eye, however, neighboring photoreceptors, bipolar cells, and ganglion cells interact with one another (Rodieck, 1998; Strauss et al., 2022). Lateral inhibition serves to differentiate and thereby increase local contrasts (Ratliff, 1965).

In fact, our eyes are constantly moving (Martinez-Conde et al., 2004a,b). Even while steadily fixating an object, retinal image positions undergo small random drifts and jitter at microscopic scales covering multiple photoreceptors, which are separated in central fovea by about 0.5 arcmin. And the image positions are interrupted at random intervals about twice per second by rapid micro-saccades of roughly 10 arcmin (Kowler, 2011; Rucci and Poletti, 2015; Intoy and Rucci, 2020). Importantly, small image motions improve spatial resolution (Rucci et al., 2007; Rucci and Poletti, 2015; Intoy and Rucci, 2020).

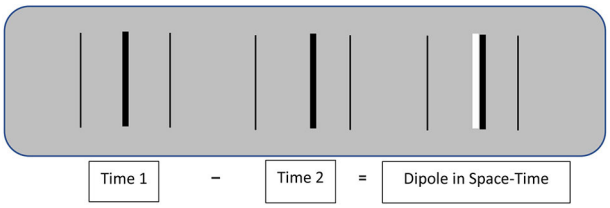


FIGURE 1
Schematic illustration of the dipole image contrast created by the image motion of a central target line. Real-time shifts in spatial position are much more visible than equivalent differences in stationary positions. The thin outside lines in this illustration are stationary reference lines.

Image motion improves spatial resolution by transforming spatial contrasts into larger contrasts in space-time. Figure 1 illustrates how this happens. A shift in the spatial position of a given image feature (a dark bar in Figure 1) produces a spatiotemporal dipole defined by the difference between successive images. The contrast of this dipole is twice that of the separate spatial images, and it has an ordered structure that specifies its displacement in both space and time. That spatial displacement is an intrinsic property of the dipole that does not involve extrinsically defined spatial positions.[1]

Intrinsic image structure

Visual information about spatial relations and motions is defined relative to a reference frame. Specifying that reference frame is fundamental for vision sciences. A common intuition is that the initial reference frame for spatial vision is the anatomical mosaic of retinal photoreceptors. Marr (1982), for example, states that "…in the case of human vision, the initial representation [of optical images] is in no doubt—it consists of arrays of image intensity values as detected by the photoreceptors in the retina" (p. 31). This statement might be taken to mean that spatial positions and relations are encoded by local signs of the photoreceptors stimulated by any given optical image. One might wonder, how else could it be?

Despite the intuitive necessity of this starting point for perceptual organization, it entails two computational problems: First, this reference frame is *extrinsic* to the spatial organization of observed objects and motions in the environment. Second, it is an implicitly static reference frame—whereas retinal positions, spatial separations, and angular directions of retinal image features are actually continually changed by movements of the observer's eyes, head, and body relative to the observed environment. A basic problem for a theory of vision is to understand how information defining objects is obtained from continually shifting images.

Representing image structure by reference to the retinal photoreceptors is consistent with our intuitive understanding of space and time as reference frames that are independent of their contents. In modern physics, space and time are derived relations among moving masses, but that abstract physical realm may seem irrelevant to visual science. Nevertheless, spatial organization can be structured by *intrinsic* spatiotemporal relations within moving images, invariant with retinal position.

Two insights about the intrinsic image information for perception are that (a) *retinal images are images of surfaces*, and (b) the 2nd-order differential structure[2] of moving images is isomorphic with that of environmental surfaces (Koenderink and van Doorn, 1975, 1980, 1991, 1992a,b; Koenderink, 1987, 1990). The interdisciplinary literature on that intrinsic image information is beyond the scope of this article, though we recently reviewed evidence about that information (Lappin and Bell, 2021). Here, we focus on the role of motion in perceptual organization.

If spatial positions and relations are represented relative to the intrinsic image structure, that does not in any way indicate the irrelevance of retinal photoreceptors and neurons. The question is whether spatial relations are anatomically defined by the positions of the receptors, or by spatiotemporal distributions of activity in the receptors and neurons.

Lappin and Craft (2000) tested alternative hypotheses about intrinsic image structure vs. extrinsic coordinates as reference frames for the optical input to vision. A first experiment tested the precision of visual information about intrinsic spatial positions in images that were rapidly and randomly jittered on the display monitor, as compared with stationary displays. The randomly jittered images disrupted extrinsically defined spatial positions on the monitor and the retina, but preserved intrinsic image structure that was invariant with the random changing positions.

Images of three lines similar to those in Figure 1 were used to evaluate visual acuities for (a) detecting motion and (b) discriminating relative positions. Two image conditions involved either *stationary* or *randomly jittered* images. In both conditions, observers used one joystick to adjust the amplitude of rapid (10/s) random displacements of the center line relative to the two flankers so that its relative motion was undetectable; and they used a second joystick to center the target line (1° in length) so as to bisect the space between the two flankers. The correct target-flanker separation was varied from 0.5° to 4°. In the randomly jittered images, the whole 3-line pattern was randomly repositioned 10 times per sec at equally probable positions in a rectangular area of 12 × 12 arcmin. The root-mean-square (RMS, standard deviation) random image displacements were 5.4 arcmin both horizontally and vertically. Figure 2 illustrates the stimulus pattern and two

1 This dipole contrast change involves a spatiotemporal relation between two images. If the position shift is defined relative to the flanking lines, then motion of the target bar changes the *relative* position of the target and flankers. Experiments have shown that such spatial relations are important for acuities in both stationary and moving images (e.g., Legge and Campbell, 1981; Lappin and Craft, 2000). Moreover, the difference between the target vs left flanker and target vs. right flanker involves a difference of differences—a change in 2nd-order differential image structure. The nature and function of such 2nd-order structure are discussed below.

2 Zero-order spatial relations are defined by absolute spatial positions; 1st-order spatial derivatives involve relations between pairs of points; and 2nd-order spatial derivatives involve relations among three points defined by differences in the pair-wise distances on either side of a given point. The 2-dimensional structure of such differences of differences identify the local shapes of smooth surfaces—as illustrated in the top panel of Figure 9.

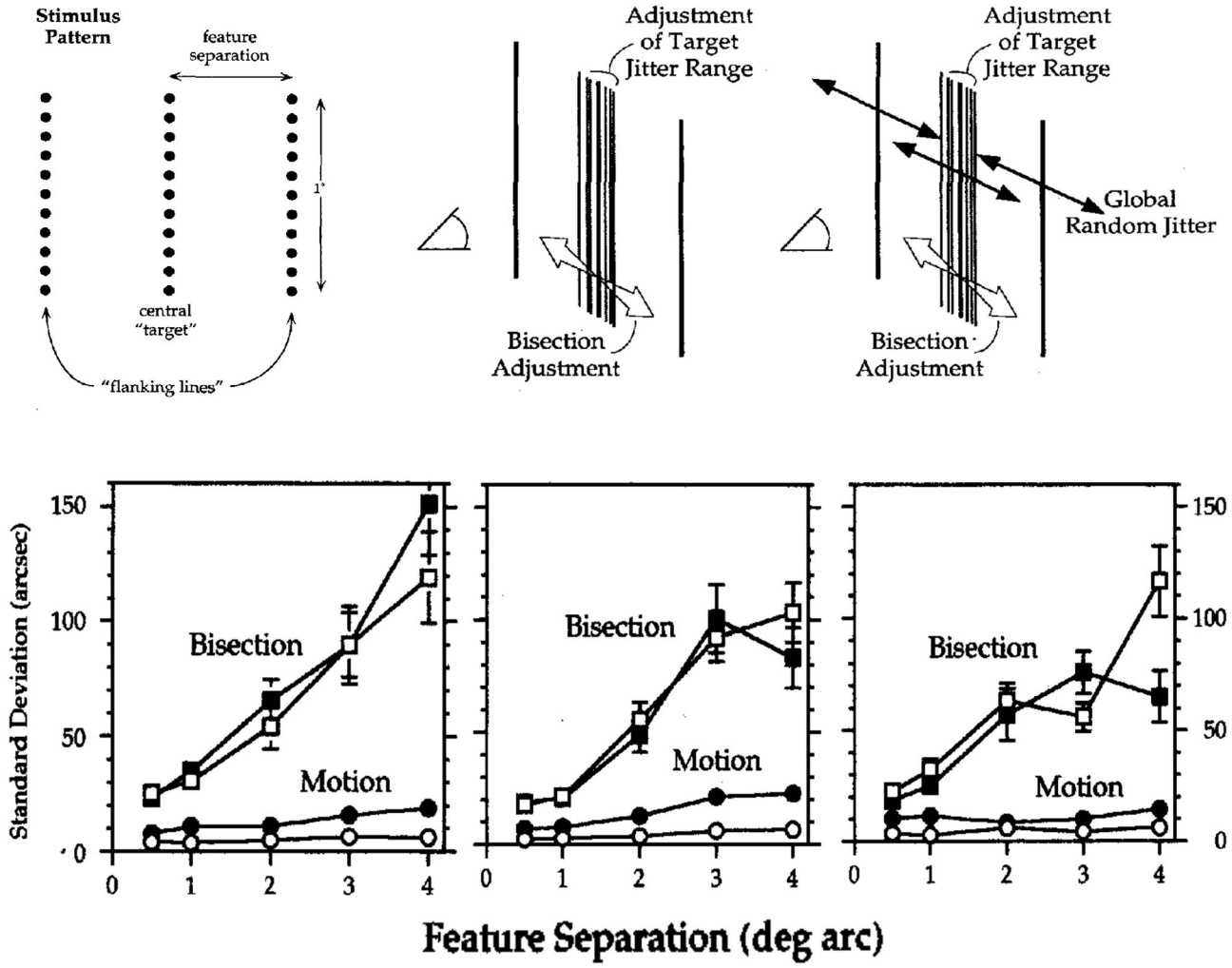


FIGURE 2
Spatial acuities for bisection and motion detection in stationary images (unfilled symbols) and images with randomly jittered global image positions (filled symbols). **(Upper)** Schematic illustrations of the visual patterns and two adjustment tasks. **(Lower)** Acuities for bisection and motion detection (standard deviations of 20 adjustments, arcsec) for three observers. Error bars are ±1 SEM; those not shown were too small to be visible (From Lappin and Craft, 2000. Copyright © 2000 by the American Psychological Association. Reproduced and adapted with permission).

concurrent adjustment tasks, along with the adjustment acuities of three observers.

The results in Figure 2 show that substantial global random image jitter had very little effect on acuities for detecting motion or bisecting spaces. Of particular interest are the "hyperacuities"[3] (Westheimer, 1975, 1979) for motion detection. In stationary images with 1° feature separations, for example, detection thresholds for three observers averaged just 3.1 arcsec—about 10% of the distance between photoreceptors in central fovea! In the randomly jittered images, with RMS horizontal jitter of 340 arcsec, average detection thresholds for relative motion increased—but to only 10.1 arcsec, less than half the half the separation between foveal cones! Thus, precise visual information about motion was based on *relative* motion—relative to the intrinsic image structure, robust over substantial random variations in retinal positions.

As expected, bisection acuities were less precise than those for motion detection, but these too were barely affected by global random image jitter. For feature separations of 1° to 4°, bisection acuities were approximately proportional to the feature separations, averaging just 0.78 and 0.83% of the separation between target and flankers.

Though not immediately evident in Figure 2, thresholds for motion detection as well as bisection increased approximately in proportion to separations of 1°-4° between the target line and flankers (Proportions were slightly greater for the 0.5° separation). For the stationary and jittered images, the motion acuities averaged just 0.06 and 0.18% of the feature separation. The proportionality of spatial resolution and feature separation also shows that perceived image motion involves intrinsic image structure, not

3 *Hyperacuity* refers to spatial resolution that exceeds a limit of about ½ arcmin or 30 cycles per degree imposed by the human eye's optical diffraction and by separations of about ½ arcmin between cones in central fovea.

local retinal positions as previously suggested (McKee et al., 1990).

Homogeneity of the visual motion field

Image motion is probably also a basis for the perceptual homogeneity of space over the whole visual field. As is well known, spatial resolution of static patterns is substantially reduced in the peripheral field. Nevertheless, despite those reduced acuities, spatial structure and motion seem subjectively constant over the visual field. That subjective phenomenology might seem surprising and puzzling—given that the densities of photoreceptors, ganglion cells, and cortical neurons all decrease rapidly with increased eccentricity, and the receptive fields of ganglion cells and cortical cells increase substantially (see Banks et al., 1991). Accordingly, spatial forms must be much larger in the periphery to be as visible as those in the fovea (Anstis, 1974). At 30 deg eccentricity, Anstis estimated that image sizes should be about 15 times larger to be seen as well as in the fovea.

Visual resolution for motion, however, is not reduced in the peripheral field. Lappin et al. (2009) evaluated spatial and temporal thresholds (which covary with motion speed) for discriminating left/right motion directions in the fovea and at ±30 deg eccentricity (High-contrast vertical gratings (1 c/deg) moved within 3 deg diameter envelopes at speeds ranging from 0.08 to 20 deg/s). Visual resolution for these motion discriminations differed sharply from that of stationary forms. For image speeds above 0.5 deg/s, spatial (and temporal) thresholds were *lower* in the periphery than in the fovea. Even at a very slow speed of 0.08 deg/s, peripheral thresholds were just 1.1 arcmin (compared to a foveal threshold about half that size). With increasing speeds, spatial displacement thresholds increased (and temporal durations decreased), but both spatial and temporal thresholds were consistently lower in the periphery. Related results were also reported by van de Grind et al. (1983, 1992). Thus, the visual motion field is more homogeneous than the visual field of static patterns.

Such homogeneity of the visual motion field was also found in experiments that evaluated perceptual relationships among multiple moving and stationary patterns—which were simultaneously presented in the fovea and at ±30 deg eccentricity (Lappin et al., 2004a,b). Using stimuli like those above (Lappin et al., 2009), *motion perception* was evaluated by temporal thresholds for direction discrimination; and *stationary form discriminations* were evaluated by orientation thresholds for discriminating stationary left/right tilts of the gratings. One experiment measured thresholds for an oddball detection task in which all three directions of motion or tilt were the same or one was different. For *moving gratings*, we found that their relative directions were easily perceived: Thresholds for the same/different motion directions were essentially the same as those for discriminating the direction of any single grating—much lower than if they were visually independent. For *stationary gratings*, however, the opposite result occurred: Thresholds for same/different tilts were much higher than those at any single location. Simultaneously perceiving stationary forms in the central and peripheral fields involved competition for attention. But the visual motion field was perceptually organized, coherent.

The visual nervous system preserves order in space-time

The importance of motion for perceptual organization is also indicated by the fidelity with which information about image motion is preserved by neural signals. Under optimal conditions, human observers can perceive the spatiotemporal order (direction) of two adjacent stimuli separated in time by only 3 ms (Westheimer and McKee, 1977). To be consciously perceived and discriminated, such differences must exist in the retina (Brindley, 1970) and remain through transmission to the cortex.

Individual retinal ganglion cells transmit information about the changing stimulation in their receptive fields by modulating their spike rates. The precision and reliability of these temporally varying signals was evaluated by Borghuis (2003) and Borghuis et al. (2019). Spike trains carrying this temporal information are illustrated in Figure 3, which shows spike trains recorded for a single retinal ganglion cell of a cat in response to moving gratings (from Borghuis et al., 2019). Dynamics of these neurons are similar for all mammals (Borghuis, 2003; Chichilnisky and Kalmar, 2003). Responses of these cells are not directionally selective, but they are highly sensitive to the temporal modulations produced by motion.

Figure 3 shows the spatial organization implicit in the temporal variations in spike rates at a given spatial position. The temporal periodicity of spike rates in the columns for the 2.0 and 8.0 Hz drift rates illustrate how an individual ganglion cell reveals the spatial structure of a sinusoidal luminance pattern moving through its receptive field. This periodic structure is robust over wide variations in contrast. Information about the direction and speed of motion is provided by phase differences in the spike trains of neighboring neurons. In short, *temporally varying spike rates carry spatial information about moving patterns.*

The temporal resolution of spike-rate variations like those in Figure 3 is limited by random variability. But this variability is reduced by temporal integration. Borghuis and colleagues (Borghuis et al., 2019) evaluated the temporal resolution of these neural signals by determining the integration times needed to distinguish stimulus-controlled spike rates from random sequences of the same inter-spike intervals. Responses to optical motions ranging from 0.5 to 16 Hz at contrasts ranging from 10 to 70% were recorded for 37 ganglion cells (33 X-type, 4 Y-type). Critical integration times were identified by cross-correlating pairs of spike trains for 20 repetitions of the same stimulus and then comparing those correlations with those for randomized sequences of the same spike trains. These cross-correlations vary as a function of the temporal interval over which the momentary spike rates are measured; and correlations were evaluated for integration times ranging from 1 to 500 ms. A well-defined peak in the difference in cross-correlations for the stimulus-controlled vs. randomized spike trains identified the optimal integration time and temporal resolution of each cell in each stimulus condition. Results are shown in panels A and B of Figure 4.

Data in the upper panels of Figure 4 show that that the temporal resolution of these retinal neurons improved rapidly as

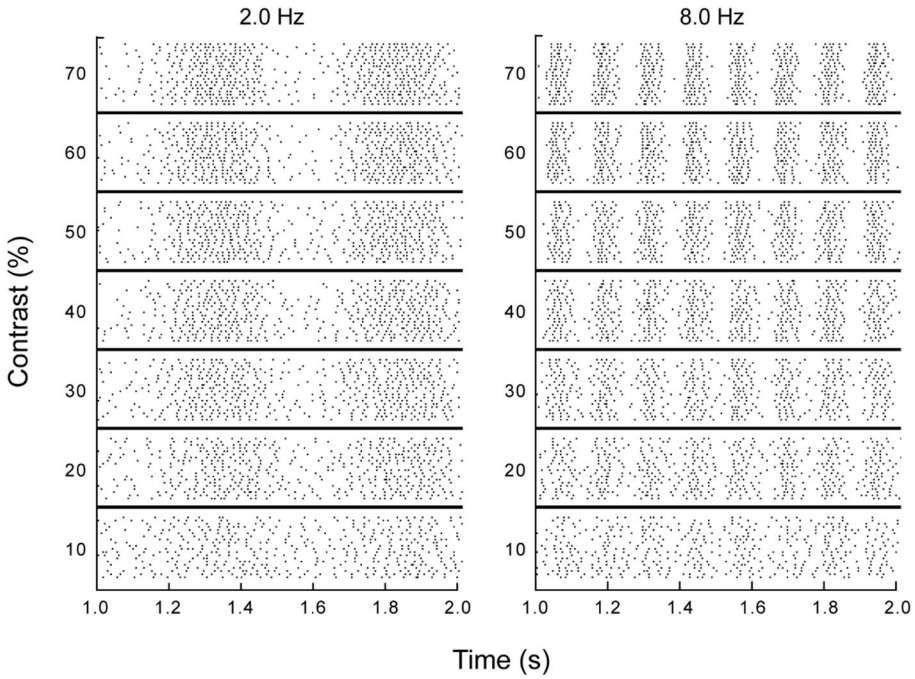


FIGURE 3
Responses of a single retinal ganglion cell to drifting sinusoidal gratings. Each row in each section is a 1 second raster plot of responses to a grating at the specified temporal frequency and contrast, with each stimulus repeated at least 20 times and each dot representing a single spike (from Borghuis et al., 2019, p. 6).

the temporal frequency of the optical oscillation increased from 0.5 to 16 Hz. Indeed, the decreases in optimal integration times were almost inversely proportional to the temporal frequency of the motion-caused oscillation. One might expect that temporal resolution would depend on the optical contrast of the gratings, but contrasts above 20% had almost no effect. The temporal resolution of these retinal signals depends on motion speed rather than contrast: Slower motions resolve more slowly.

How does the temporal resolution of retinal ganglion cells relate to discriminations of motion direction—which depend on *phase differences* among the spike rates of neighboring cells? Borghuis et al. (2019) addressed that question by evaluating temporal duration thresholds for human observers' discriminations of left/right motion directions for similar moving gratings. Stimuli for the human observers were Gabor patches: 0.33 deg diameter at $\pm 2\sigma$ width of a Gaussian envelope, 3.0 c/deg spatial frequency, with temporal frequencies from 0.5 to 32 Hz and contrasts from 5 to 80%.

Results are shown in panel C of Figure 4. Remarkably, these two aspects of information about moving visual images—one involving single retinal cells, and the other involving large numbers of cortical cells—were both qualitatively and quantitatively similar. The correlation between human duration thresholds and neural time constants was $r = 0.99$ for retinal X cells, and $r = 0.98$ for retinal Y cells. For fast motions at 16 Hz, the integration time constants for both X and Y retinal cells were slightly lower than the human discrimination thresholds; but for very slow motions at 0.5 Hz, requiring temporal integration approaching 100 ms, human discriminations exhibited slightly better resolution than the (cat's) retinal cells (This advantage for the humans probably reflects their greater spatial acuity).

Similar temporal resolution of motion by retinal neurons and human discriminations of motion directions, involving phase relations among multiple neurons, implies that the visual system preserves the spatiotemporal order of moving images with very little information loss from retina to cortex.

Coherent phase relations among spatially separate motion signals

How does vision resolve *different* image motions? Image motions are produced by the eyes and body as well as external objects. How does vision differentiate relative motions? In fact, vision is extraordinarily sensitive to relative motion. And the spatial resolution of relative motion implies correlated—*coherent*—phase relations among spatially separate retinal signals.[4]

Lappin et al. (2001) evaluated visual resolution of relative motion by measuring observers' abilities to detect phase differences in sinusoidal oscillations of spatially separate image features. The stimuli were three horizontally aligned and horizontally oscillating Gaussian luminance blobs [The apparent diameter of the blobs was roughly 1/2 deg ($\sigma = 7.1$ arcmin) and the peak luminance

4 "Coherence" is used here with essentially the same meaning as in optics and lasers—based on and measured by correlated variations. In both optics and vision, coherent phase relations resolve spatial scales much finer than the wavelengths of uncorrelated variations.

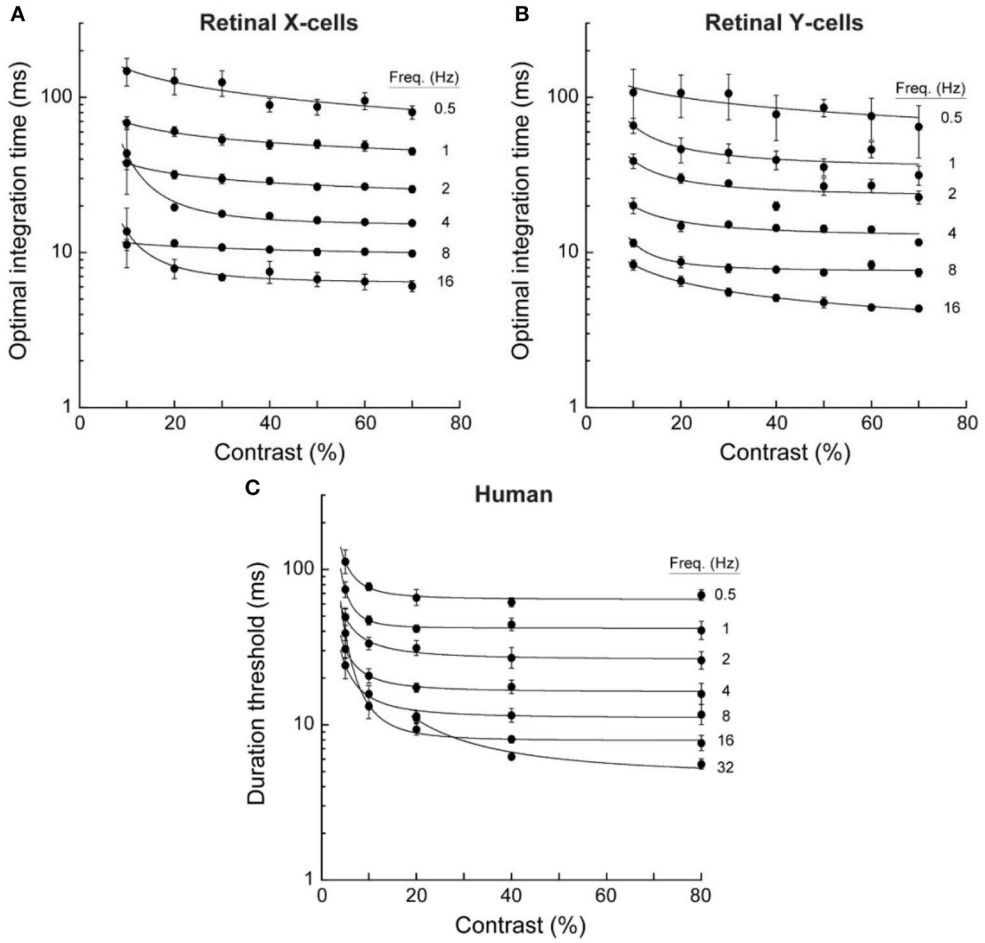


FIGURE 4
Optimal integration times for neural responses of retinal ganglion cells and duration thresholds for human direction discriminations. Data were averaged for 33 retinal X cells, 4 retinal Y cells, and 4 human observers. Error bars are for SEM. These temporal integration durations decreased with temporal frequency but were little affected by contrast above about 10% (adapted from Borghuis et al., 2019, p. 9).

was 78% above the background]. In one experiment with 100 arcmin center-to-center separation between blobs and 1.5 Hz sinusoidal oscillations, average threshold acuities (at $d' = 1.0$) for discriminating in-phase vs. anti-phase motions of the central blob relative to the two flankers were *lower* than those for detecting rigid motions of all three blobs—8.7 vs. 11.0 arcsec. This hyperacuity for relative motion involved image displacements of only 0.14 and 0.18 pixels (on a monitor with 1,024 pixels horizontal resolution) and 0.14 and 0.18% of the space between blobs.

Because these perceived relative image displacements were fractions of single pixels and fractions of separations between foveal photoreceptors, the visual temporal variations produced by individual pixels as well as retinal photoreceptors and neurons must be correlated. This spatial hyperacuity derives from temporal phase differences. Indeed, discrimination thresholds for spatial oscillations provide an estimate of the correlation between visual oscillations of the center and flanking features: Thresholds, at $d' = 1.0$, estimate standard deviations corresponding to visually detected motion distances. By a geometric construction, where the cosine of the angle between two vectors equals their product-moment correlation, the law of cosines gives an estimate of the correlation.[5] If 11.0 arcsec is the distance of the in-phase center and flanker motions, and if 8.7 arcsec is the anti-phase motion distance, then $r = 0.67$. Coherent retinal signals are necessary for the obtained hyperacuities.

A similar experiment evaluated acuity by varying the relative oscillation phase of the central blob. For Gaussian blobs separated by 320 arcmin and a 1.6 arcmin oscillation, the threshold phase difference was $<18°$. This acuity was robust over increased spatial separations. For separations of 80, 160, and 320 arcmin, threshold values were 0.24, 0.36, and 0.49 arcmin (0.3, 0.2, and 0.15% of the separation). The acuities were also robust over varying temporal frequencies—best at 3 Hz and increasing from about 0.25 arcmin to 0.5 arcmin at 9 Hz.

5 Let S represent the standard deviations corresponding to the motion thresholds at $d' = 1.0$. And let the subscripts C, F, and D designate vectors corresponding to motion distances of the center, flankers, and center-flanker difference—where D is the difference vector between the ends of C and F joined at their base. By the law of cosines, $S_D^2 = S_C^2 + S_F^2 - 2\ S_C\ S_F\ (\cos\ \delta)$, where δ is the angle between vectors C and F, and $\cos\ \delta = r$. Lappin et al. (2001) used a more complex formulation that yielded a similar value.

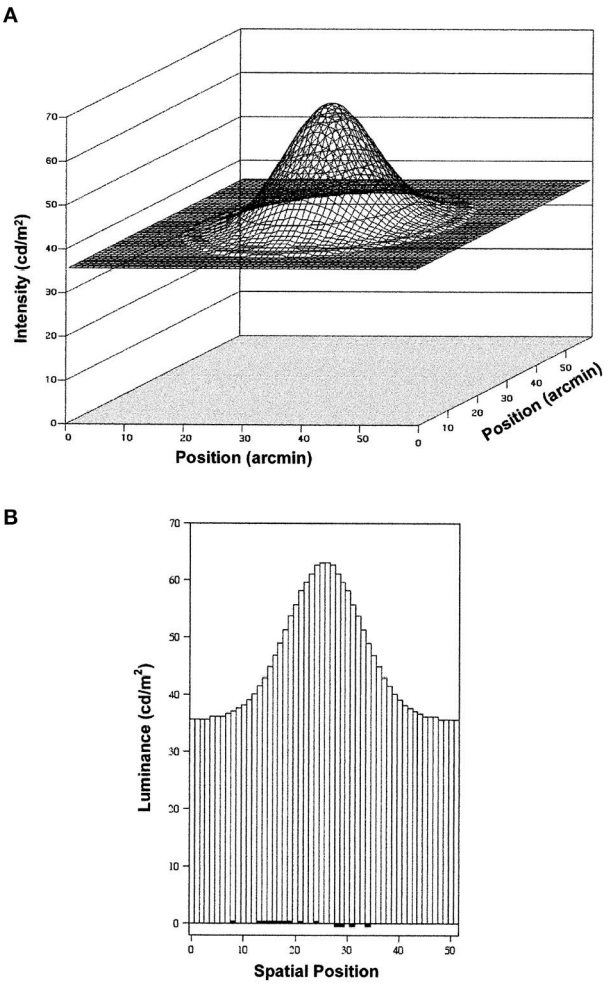


FIGURE 5
Luminance distributions for the Gaussian blobs and background in the study of Lappin et al. (2001). **(A)** is a smoothed surface plot of the luminance distribution, and **(B)** is a histogram showing the luminance at each 1 arcmin^2 square pixel at a cross-section through the center of the blob. Dark areas at the bottom of this histogram show the increased and decreased luminance produced by a detectable leftward shift of 0.14 pixel. That displacement changed the luminance by one gray-scale unit (8-bit resolution, ~0.5 cd/m^2) in 20% of the 1,272 pixels in the blob (reprinted from Lappin et al., 2001, p. 718).

The visual precision in this study is also remarkable when evaluated by the tiny changes in luminance (8-bit grayscale resolution) produced by small image motions. A lateral shift of 0.14 pixels produced no luminance change at either the center or outside edges of the blob, nor indeed any luminance change in 80% of the blob's pixels. These detectable changes in relative image positions involved a change of about 1% of the initial luminance of just 20% of the 1,272 pixels in the Gaussian blob (The area of each pixel was 1 arcmin^2, stimulating several neighboring photoreceptors, and each blob stimulated several thousand photoreceptors). Relative to the total blob luminance, the contrast change produced by a shift of 0.14 pixels was just 0.24%. Figure 5 illustrates the luminance distribution and the dipole change produced by a leftward shift of 0.14 pixels. Visual sensitivity to such tiny changes in the spatiotemporal distribution of stimulation entails the correlated responses of many thousands of adjacent and separated retinal receptors and neurons.

The dipole structure of these motion-produced changes in stimulation is important for the visual sensitivity to motion. The visual importance of this dipole structure is shown by the results of experiments by Lappin et al. (2002). Discriminations of relative motion were compared with those for equivalent contrast changes in symmetrical (not dipole) oscillations that do not alter the blob's spatial position. Single blobs ranging in size from $\sigma = 3$ to 60 arcmin were oscillated at 3 Hz. Thresholds for detecting the stationary symmetrical oscillations were about 3 times greater than those for oscillating motions, although both were similar for the smallest blobs. The detection thresholds for contrast changes in large blobs averaged 0.09% for motion but 0.23% for stationary oscillations.

Lappin et al. (2002) evaluated the perceptual organization of these moving vs. stationary image changes by testing discriminations of *phase differences* in oscillations of center and flanking blobs. As expected, phase differences in *motion* were visually salient and effortlessly discriminated, but phase differences in the *stationary symmetrical* oscillations were difficult to perceive even when contrast oscillations of individual blobs were large and easily visible. Averaged across spatial separations of 80 and 240 arcmin and oscillation frequencies of 1, 3, and 8.5 Hz, contrast thresholds for discriminating in-phase vs. anti-phase oscillations averaged 0.21% for relative motion and 1.38% for stationary contrast oscillations. Thus, image motions were visually coherent, but stationary contrast oscillations were not.

As described above (footnote 5), correlations between these visual signals can be estimated from the oscillation detection thresholds for the center blob by itself, the two flanking blobs alone, and in-phase vs. anti-phase oscillations of the central and flanking blobs. These threshold estimates were obtained for three Gaussian blobs ($\sigma = 10$ arcmin) oscillating at 3 Hz, and separated in the phase discrimination task by 100 arcmin, and by 200 arcmin between the two blobs in the flanker oscillation threshold task. Thresholds for detecting oscillations of the center, flankers, and center/flanker phase difference were, respectively, 0.33, 0.34, and 0.23 arcmin. The estimated correlation was $r = 0.76$.

For the stationary (symmetrical) contrast oscillations, however, the thresholds (in corresponding spatial values) averaged 0.50, 0.69, and 1.32 arcmin, yielding an estimated negative correlation beyond $r = -1.0$ (see Lappin et al., 2002, for the computational rationale). Without very large contrast changes, the relative contrast oscillations of the separated image features were not simultaneously perceived.

Statistical coherence of perceived structure from motion

To perceive the organized structure of images, the visual system must integrate *common motion*. Aspects of the process resemble auto-correlation—a linear statistical correlation between optical patterns at neighboring spatial and temporal locations (Reichardt,

1961; Uttal, 1975). Auto-correlation functions are defined on the *transformations* that map successive images onto one another, not on the image coordinates *per se*. The general form of such functions can be written as

$$A(\varphi) = \iint \varphi\left[f_1(x, y)\right] \cdot f_2(x, y)\, dx\, dy$$

where φ is a transformation that maps a 2-dimensional image $f_1(x, y)$ at time 1 onto image $f_2(x, y)$ at time 2. Auto-correlations are sometimes presented as functions of horizontal and vertical translations, $\varphi[f(x, y)] = f(x + \Delta x, y + \Delta y)$. In that special case, the functional distinction between the (x, y) image coordinates and the motion parameters Δx and Δy is not obvious. In more general cases, however, moving objects and observers often change relative viewing directions (rotating in 3D), which changes relative spaces between neighboring image features. And autocorrelations can also be defined on such image transformations. Psychophysical experiments have tested applications of autocorrelation to the statistical characteristics of both 2D and 3D structure from motion.

The statistical nature of motion integration was evident in early studies with random-dot patterns. Direction discriminations were found to increase with the area, number of elements, and inter-frame correlation, and decrease at low contrast and with greater spatial and temporal separations between images (e.g., Bell and Lappin, 1973; Lappin and Bell, 1976; van Doorn and Koenderink, 1982a,b; Chang and Julesz, 1983; Williams and Sekuler, 1984; van Doorn et al., 1985; van de Grind et al., 1992). To a reasonable approximation, visual detection of the coherence of these patterns operates as a linear system: The output signal/noise ratio (d') of motion discriminations increases proportionally with the input signal/noise ratio—with the percentage of elements with the same displacements, the square root of the number of elements, and the square root of the number of frames (Lappin and Bell, 1976; Lappin and Kottas, 1981).

The visual coherence of moving images is not limited to 2D translations. Rotations in the image plane, for example, involve a 360° range of directions and velocities that increase from the center of rotation. Nevertheless, discriminations of rotation direction and statistical coherence are as accurate as those for 2D translations (Bell and Lappin, 1979; Lappin et al., 1991).

Lappin et al. (1991) found similar visual sensitivities to the statistical coherence of barely visible small rapid *random* translations, rotations, expansions/contractions, and combinations of those transformations. In one experiment, the image positions of sparse dot patterns (e.g., 8 equally spaced dots on the circumference of a 10 deg diameter circle) were randomly sampled from normal distributions at 50 Hz for 1s, and observers discriminated between coherent images in which all dots were displaced by the same transformation vs. those in which displacements were independent for each dot. If observers could see any motions at all, they could see whether they were coherent or incoherent. And discrimination thresholds were similar for each transformation.

A similar experiment, also with small 50 Hz random image transformations, tested perceptual interactions between transformations. Coherence discriminations for one transformation were evaluated alone or when added to coherent random changes produced by another transformation. Coherence detections of rotations and expansions were the same whether or not one was added to the other. Violations of such linear independence were found when rotations or expansions were combined with random horizontal and vertical translations, but these violations were not large. Discriminations between coherent and incoherent rotations were 95% correct even when added to noisy backgrounds of uncorrelated translations of each dot. Thus, the perceptual organization of these rapid random image changes was governed by an essentially linear visual representation of the whole spatial pattern.

Importantly, the limiting spatial parameters for detecting these coherent motions are defined on the image rather than the retina. Limiting displacement distances between frames are proportional to the size of the pattern rather than the retinal distance (Bell and Lappin, 1973, 1979; Lappin and Bell, 1976; Chang and Julesz, 1983). This image scale-invariance is contrary to the idea that the perception of these patterns involves a "short-range process" limited by retinal spacing (Braddick, 1974).

Nonlinear visual coherence of three-dimensional structure and motion

The perceptual role of spatiotemporal structure is also clear in experiments on the perception of 3D structure from motion (e.g., Johansson, 1973; Braunstein, 1976; Rogers and Graham, 1979; Lappin et al., 1980; Doner et al., 1984; Todd and Norman, 1991; Perotti et al., 1998; Lappin and Craft, 2000). Analogous to Julesz's demonstrations of "cyclopean" perception of random-dot stereograms (Julsez, 1971; Julesz and Tyler, 1976; Tyler and Julesz, 1978), similar perception of surfaces moving in depth can also be achieved with two frames of random-dot cinematograms—where the perceived 3D organization derives from visually coherent motion between frames but is invisible in either frame alone.

Unlike the approximate linearity of perceived 2D image dynamics, however, perception of 3D structure and motion evidently involves nonlinear organization. Smooth surface structure and coherent motion between the successive frames are found to be important for perceiving the 3D organization.

As illustrated in Figure 6, Lappin et al. (1980) displayed two frames of dots randomly positioned on the surface of a sphere rotated around its central vertical axis between frames, with each frame 200 ms and no inter-frame interval. Despite each dot shifting in a curved trajectory that varies with its spherical position and in opposite directions on the front and back surfaces, the smooth surface and motion are immediately obvious to most naïve observers—*if* the dot positions are perfectly correlated in the two frames of the rotated surface.

The accuracy of these perceptions was evaluated by observers' coherence discriminations for patterns with different inter-frame correlations. Observers were about 63% correct in discriminating between displays with 100 vs. 94% correlated dot positions, and more than 80% correct in discriminating 100 vs. 75% correlated

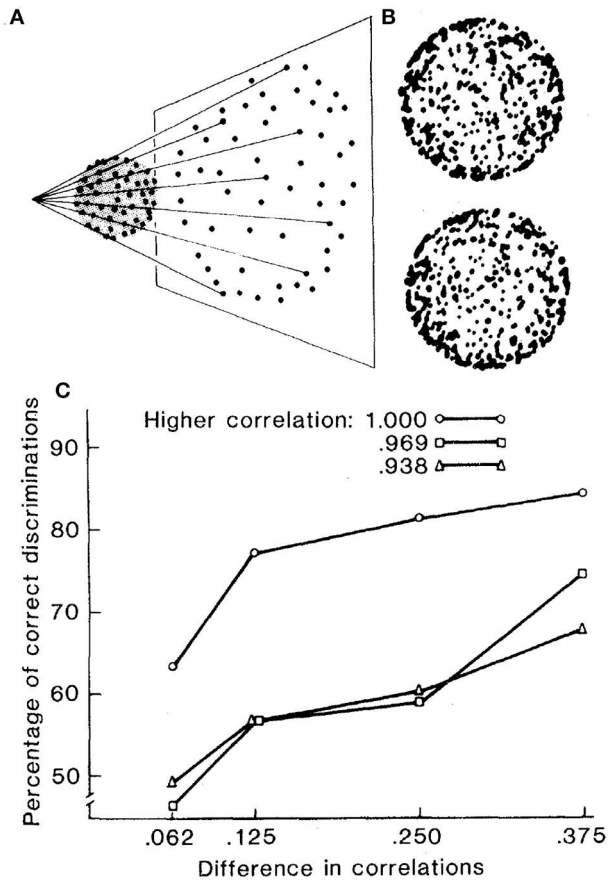


FIGURE 6
(A) Schematic illustration of the projected image of 512 dots randomly distributed with uniform probability density over the surface of a transparent sphere. **(B)** Photographic images of two successive frames of the sphere rotated by 5.6° around its vertical axis. Each frame was displayed in the experiment for 200 ms, with no interval between frames. **(C)** Average accuracies of four observers in discriminating between patterns with two different correlations in a 2-alternative forced choice task (reprinted from Lappin et al., 1980, p. 718).

patterns. But small reductions in inter-frame correlations disrupted the visual coherence. Accuracy declined to <60% in discriminating patterns with 97 vs. 72% correlations. Thus, the visual coherence was nonlinear.

This nonlinear stability was studied further by Doner et al. (1984). Coherence discriminations improved substantially with added frames, and also when frame durations were reduced from about 240 to 60 ms. These and other results indicate that visual coherence of this 3D structure and motion involves self-organizing, globally cooperative processes. The coherent organization required integration of signals from opposite motion directions of dots on the transparent front and back surfaces of the sphere plus smoothly varying displacements in images of the spherically curved surfaces. Even though this 3D structure and motion is quite perceivable, its structural coherence is evidently less immediate than that of 2D image motions. Perceptual organization of some moving patterns seems to entail nonlinear visual dynamics (see Strogatz, 2003, 2018).

Spatial forms defined by differential motion

The statistical nature of visual coherence in dynamic random dot patterns might suggest that visual organization is mainly integrative, gaining information from common motion. Visual integration is insufficient, however. The spatial structure of moving optical patterns also involves differential motion, and perceiving that structure requires spatial differentiation as well as integration. Integration and differentiation are opposed but basic inter-dependent aspects of visual organization.

Evidence about a basic visual mechanism for differentiating image motion comes from discoveries by Tadin and colleagues. Tadin et al. (2003) found converging evidence about a counter-intuitive phenomenon: *Larger moving patterns are often less visible*. Evidently, visual motion mechanisms involve *spatial suppression*.

Figure 7, from Tadin et al. (2003), illustrates the suppressive effects—measured by temporal thresholds for discriminating the directions of gratings (1 cycle/deg) drifting (2°/s) within stationary Gabor patches of varied size and contrast. The motion directions of high-contrast gratings became substantially less discriminable as size increased from 0.7° to 5°; and large patches became less discriminable as contrast increased from 2.8 to 92%. The size of the most discriminable motions decreases as contrast increases (Tadin and Lappin, 2005). Tadin (2015) reviews many of the findings that clarify both neural mechanisms and visual functions of this spatial suppression. The neural mechanism—for integrating small patterns with low contrast and suppressing large patterns with high contrast—involves the center-surround antagonism of receptive fields of motion sensitive neurons in cortical area MT (Pack et al., 2005; Tadin et al., 2011; Tadin, 2015).

Center-surround antagonism is widespread in the visual nervous system because optical information entails spatiotemporal variations rather than merely total energy. This neural antagonism adds information—by *segregating figure from ground*.

Converging causal and correlational evidence about the role of spatial suppression in figure/ground organization was described recently by Tadin et al. (2019). Causal evidence was provided by the opposite effects of luminous contrast on two aspects of perceptual organization of moving patterns. A *form discrimination* task measured duration thresholds for discriminating the shape of an embedded form defined by opposite motion directions inside and outside the form. A *motion discrimination* task measured duration thresholds for discriminating the background motion without an embedded form. The two tasks are illustrated in Figures 8A, B.

Figure 8C shows that the contrast of these moving patterns had reciprocal effects on the time durations needed for these two discrimination tasks. Increased contrast multiplied the duration thresholds for discriminating the background motion direction; but the same increased contrast divided the duration thresholds for form discrimination by an almost equal amount. Spatial suppression was responsible for both effects. Increased contrast suppressed the spatial integration of motion signals needed to discriminate motion directions, but this same suppression enhanced the spatial differentiation of motion in the form discrimination task.

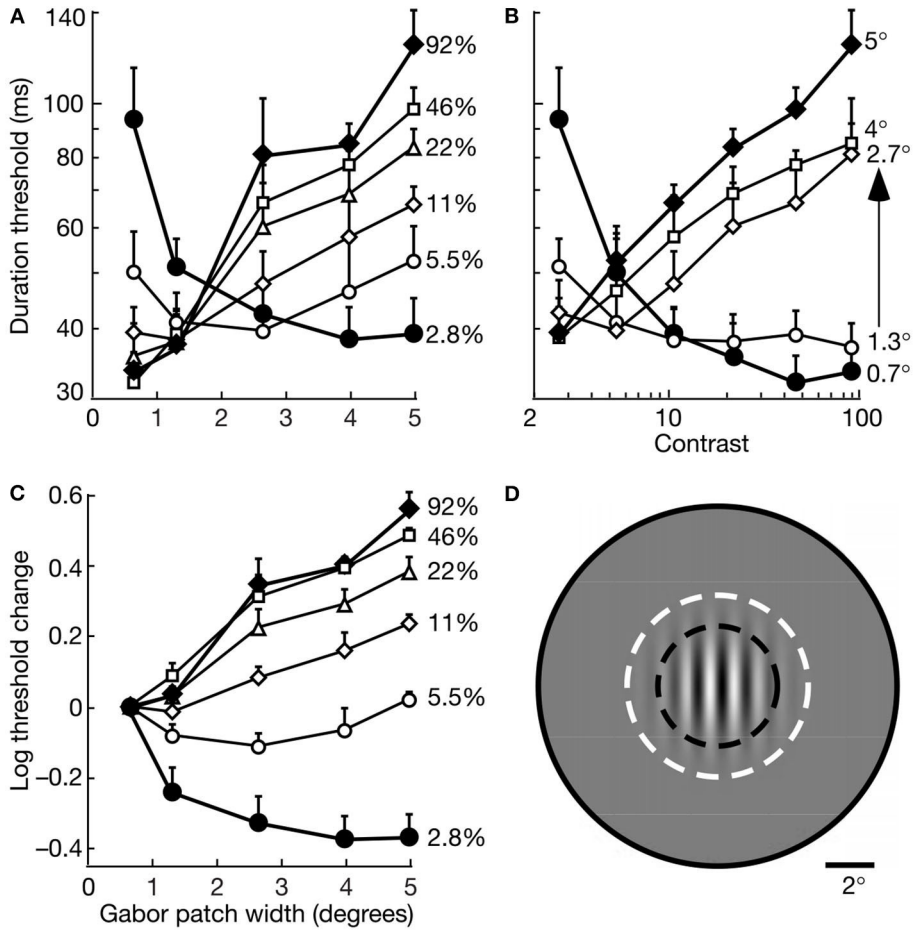


FIGURE 7
Motion discrimination depends on interactive effects of size and contrast. Data points are average thresholds for five observers. **(A)** Duration thresholds as a function of size for varied contrasts. **(B)** Duration thresholds as a function of contrast for varied sizes. **(C)** Log_{10} of threshold change as a function of size at varied contrasts. For each observer, the threshold change was evaluated relative to the duration threshold at the smallest size (0.7°) at each contrast. **(D)** A Gabor patch 2.7° wide is shown relative an average macaque foveal MT receptive field. The dark dashed lines indicate the size at which such cells often exhibit surround suppression, and the size of the surround is indicated by the full gray circle. The white dashed lines indicate the ±3σ radius of the Gabor patch (reprinted from Tadin et al., 2003, p. 313; Figure 1).

The perceptual effects of this suppressive mechanism are also found in its influence on the perceptual characteristics of several different observer populations (Tadin, 2015). Older observers, for example, are found to be better (lower duration thresholds) than control populations in discriminating large high-contrast motions (e.g., Betts et al., 2009). These effects have been shown to be linked to reduced spatial suppression.

Importantly, the better perception of large high-contrast motion patterns by older observers is accompanied by a reduced ability to perceive spatial forms defined by differential motion. Interactive effects of age on integrating and differentiating moving patterns were demonstrated by Tadin et al. (2019). Form discrimination and motion discrimination by younger and older observers were evaluated for small patterns as well as large patterns like those in Figures 8A, B. As found in previous studies, the older observers had less spatial suppression and better motion discrimination of large patterns; but the older observers also had reduced efficiency in segregating forms defined by differential motion. This interactive effect of age on motion discrimination and form discrimination is correlational evidence that visual integration and differentiation play reciprocal roles in perceiving moving images.

Perception of surface structure from moving images

As mentioned earlier in this article, two basic principles of visual perception are that (a) retinal images are primarily images of surfaces, and (b) the 2nd-order differential structure of spatiotemporal images is isomorphic with that of environmental surfaces. These insights are mainly from Koenderink and van Doorn (1975, 1980, 1991, 1992a,b), Koenderink (1987, 1990). Local surface shape is quantified by the relative values of minimal and maximal curvature, and is specified in images by spatial changes produced by rotation in depth or by stereoscopic views. A substantial literature of theoretical, psychophysical, and engineering research has validated this spatiotemporal image

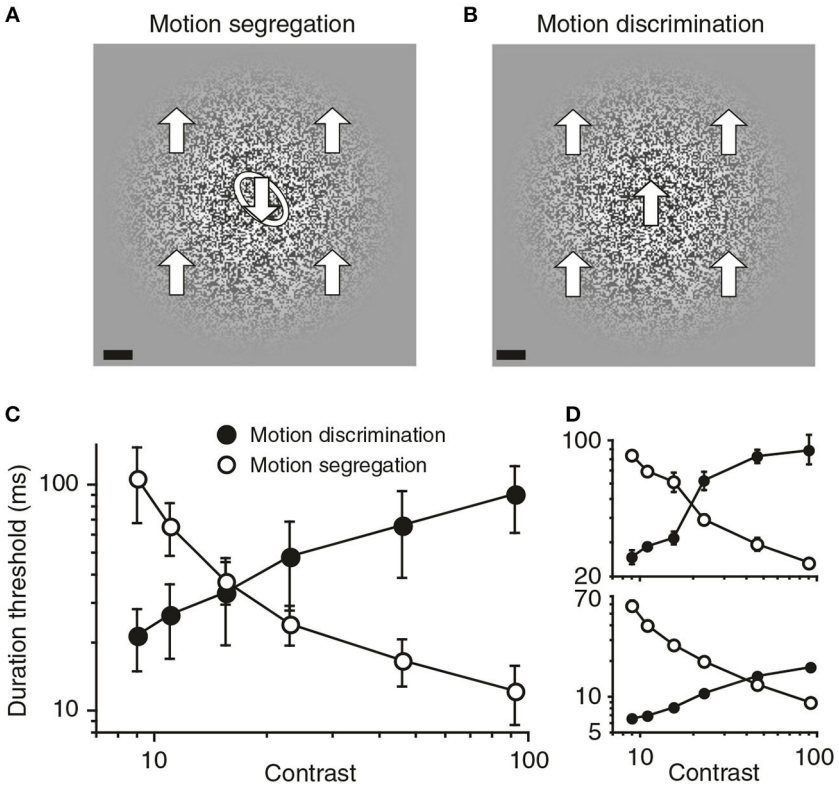


FIGURE 8
Visual segregation of an embedded form defined by differential motion becomes more effective when the background motion is visually suppressed. **(A)** In the motion segregation task, observers identified the tilt of a motion-defined oval, which could be tilted either left (as shown) or right. The large arrows that indicate motion directions and the white outline of the embedded oval are added here only for purposes of illustration. The scale bar at bottom left is 1°. **(B)** In the motion discrimination task, the background in **(A)** was presented without the embedded oval, and observers discriminated up vs. down directions of motion. **(C)** Group data showing the opposite effects of stimulus contrast on motion discrimination and motion segregation. Error bars are SEM. **(D)** Data for two individual observers (reprinted from Tadin et al., 2019, Figure 1, p. 3).

information about surface shape. The top panel of Figure 9 illustrates the geometrical correspondence between 2nd-order structures of image motions and surface shapes.

Perception of this image information about surface shape was tested by Lappin and Craft (2000). The experimental strategy was analogous to that in Figure 2: We quantified visual acuities for the 3D position of a point on curved and planar surfaces, and tested the invariance of that perceptual precision under perturbations of lower-order spatial structure. Acuity for relative 3D position was tested with image motions produced both by rotating surfaces in depth and by stereoscopic images.

The spatial patterns were hexagonal arrays of 19 dots orthographically projected onto either a spherical or a planar surface, as illustrated in the lower left panel of Figure 9. The 19 dots were equally spaced on the surfaces, but their relative image positions were changed by tilting the surfaces in depth, by 20° around both horizontal and vertical axes, so that the surface normal was slightly upward to the left. The patterns were also varied by random image rotations of 0° to 50° around the direction of view (the lines and speckled texture in this figure are for illustration only and were not in the 19-dot patterns seen by the observers). The average image separation between adjacent dots was 1°. In the moving images, surface shapes were produced by two-frame alternating ±3° rotations around the vertical axis. In the stereoscopic images, the binocular disparity between the central target dot and its nearest neighbors was 5.9 arcmin. The observer's task in both moving and stereoscopic displays of both spherical and planar surfaces was to adjust the center dot to be equidistant from the 6 surrounding dots. In the images of the spherical surface, this central surface position was neither collinear with nor centered between the surrounding dots. That position on the spherical surface was defined by the relative curvatures in two directions. The relative values of the minimal and maximal curvatures specify the surface *shape* but not its extension in depth. Image C in the lower left panel of Figure 9 illustrates how spherical shape is specified by these relative image curvatures.

For the planar surface, however, that central surface position was specified by both collinearity and bisection in any single image of the surface. Collinearity and bisection are both 2nd-order relations among three points in a single image direction. As illustrated in image B in the lower left panel of Figure 9, images of the tilted planar surface preserved both collinearity and bisection. To isolate these (1-dimensional) 2nd-order relations on the planar surface, 1st-order (pair-wise) information about spatial separations was disrupted by random image expansions/contractions (±2%) that were uncorrelated between the two stereoscopic images and

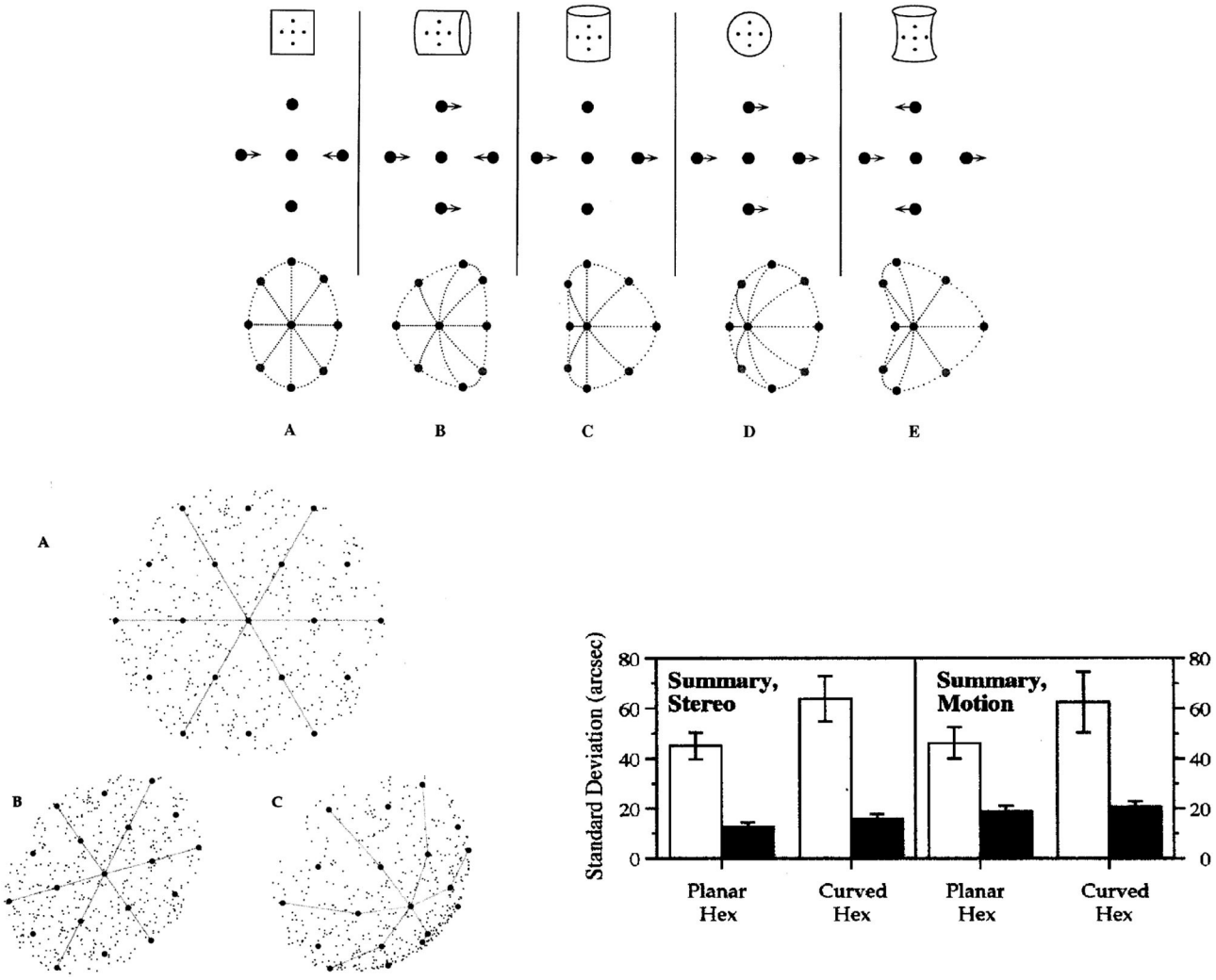


FIGURE 9
(Top) Five qualitatively different surface shapes are specified by their relative curvatures in the two orthogonal directions of minimum and maximum curvature. These five categories of surface shape are identifiable by 2nd-order image differences in stereoscopic images and relative motions produced by rotating the surface in depth. Relative values of the two principal surface curvatures are specified by such image pairs, but curvedness of the surface in depth is ambiguous. **(Lower left)** A schematic illustration of the planar and spherical surfaces used to test acuities for the position of the central target dot in 19-dot hexagonal patterns. The lines and speckled textures are added here only for purposes of illustration, and were not in the 19-dot patterns seen by the observers. The patterns contained 19 dots equally spaced in a hexagonal array, as seen in pattern A. Pattern B is an image of the planar surface tilted down and to the right. Pattern C is an image of the equally tilted spherical surface. Collinearities and relative spaces between dots are changed in the image when the spherical surface is tilted in depth, but neither bisection nor collinearity is altered by tilting the planar surface. **(Lower right)** The average acuities of three observers for centering surface position of the target dot (involving bisection and collinearity) and for adjusting the depth of the target dot onto the surface. Light bars indicate spatial bisection acuities, and dark bars indicate thresholds for binocular disparity or relative motion between the surface and target point (From Lappin and Craft, 2000. Copyright ©2000 by the American Psychological Association. Reproduced and adapted with permission).

uncorrelated between the two successive images in the relative motion displays.

The observers used two joysticks to adjust two aspects of the 3D spatial position of the target dot at the center of the 19-point hexagon—to bisect the surface space between the surrounding dots, and to position it in depth onto the surface. The latter judgments involved minimizing relative motion or stereo disparity between the target point and the smooth surface specified by the other 18 points. Importantly, the latter judgments involved the *shape* the surface, not its curvedness or slant in depth, and not a depth scale *per se*.

The average stereoscopic and relative motion acuities of three observers for the bisection and depth positions on planar and spherical surfaces are shown in the lower right panel of Figure 9. The findings of principal interest are the hyperacuities for 2nd-order image information about surface shape. The average stereoacuities were 12.1 arcsec for the planar surface and 15.4 arcsec for the spherical surface. The average acuities for relative motion were 18.6 arcsec for the planar surface and 20.4 arcsec for the spherical surface. Subjectively, the perceived surface shapes were clear and unambiguous, consistent with the obtained hyperacuities.

Not surprisingly, bisections of relative distances on these surfaces were much less precise than placing the point's depth relative to the surface. And bisecting spaces were less precise on the curved than on the planar surface—presumably because the depth scale of the surfaces, rather than their shapes, is optically ambiguous in both stereoscopic and relative motion patterns.

General principles

The psychophysical and physiological results reviewed in this article indicate that the visual system obtains precise information about environmental surface structure from the intrinsic spatiotemporal structure of moving images. The eye's remarkable spatial acuity and contrast sensitivity derive from (a) dipole image contrast changes produced by image motions, (b) coherent temporal phase relations among spatially distributed neural response patterns, and (c) preservation of this spatiotemporal structure from retina to cortex. Spatiotemporal image structure is preserved to a surprising extent in transmission through the visual nervous system. In short, perceptual organization derives from the visual coherence of moving images.

Author contributions

This article was written by both authors, and both approved the submitted version. JL was the primary author, and HB contributed editorial suggestions, scientific content, and references.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Anstis, S. M. (1974). A chart demonstrating variations in acuity with retinal position. *Vision Res.* 14, 589–592. doi: 10.1016/0042-6989(74)90049-2

Banks, M. S., Sekuler, A. B., and Anderson, S. J. (1991). Peripheral spatial vision: limits imposed by optics, photoreceptors, and receptor pooling. *J. Opt. Soc. Am. A* 8, 1775–1787. doi: 10.1364/JOSAA.8.001775

Bell, H. H., and Lappin, J. S. (1973). Sufficient conditions for the discrimination of motion. *Percept. Psychophys.* 14, 45–50. doi: 10.3758/BF03198616

Bell, H. H., and Lappin, J. S. (1979). The detection of rotation in random-dot patterns. *Percept. Psychophys.* 26, 415–417. doi: 10.3758/BF03204169

Betts, L. R., Sekuler, A. B., and Bennett, P. J. (2009). Spatial characteristics of center-surround antagonism in younger and older adults. *J. Vision* 9, 1–15. doi: 10.1167/9.1.25

Borghuis, B. G. (2003). *Spike timing precision in the visual front-end* (Ph.D. thesis). Utrecht University Library, Utrecht, The Netherlands. ISBN: 90-393-3293-2.

Borghuis, B. G., Tadin, D., Lankheet, M. J. M., Lappin, J. S., and van de Grind, W. (2019). Temporal limits of visual motion processing: Psychophysics and neurophysiology. *Vision* 3, 1–17. doi: 10.3390/vision3010005

Braddick, O. (1974). A short-range process in apparent motion. *Vision Res.* 14, 519–527. doi: 10.1016/0042-6989(74)90041-8

Braunstein, M. L. (1976). *Depth perception through motion*. New York: Academic Press. doi: 10.1016/B978-0-12-127950-9.50010-1

Brindley, G. (1970). *Physiology of the Retina and Visual Pathway*. Baltimore: Williams and Wilkens.

Chang, J. J., and Julesz, B. (1983). Displacement limits, directional anisotropy and direction versus form discrimination in random-dot cinematograms. *Vision Res.* 23, 639–646. doi: 10.1016/0042-6989(83)90070-6

Chichilnisky, E. J., and Kalmar, R. S. (2003). Temporal resolution of ensemble visual motion signals in primate retina. *J. Neurophysiol.* 23, 6681–6689. doi: 10.1523/JNEUROSCI.23-17-06681.2003

Doner, J., Lappin, J. S., and Perfetto, G. (1984). Detection of three-dimensional structure in moving optical patterns. *J. Exper. Psychol.* 10, 1–11. doi: 10.1037/0096-1523.10.1.1

Intoy, J., and Rucci, M. (2020). Finely tuned eye movements enhance visual acuity. *Nat. Commun.* 11, 795. doi: 10.1038/s41467-020-14616-2

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* 14, 201–211. doi: 10.3758/BF03212378

Julesz, B., and Tyler, C. W. (1976). Neurontropy, an entropy-like measure of neural correlation in binocular fusion and rivalry. *Biol. Cyber.* 23, 25–32. doi: 10.1007/BF00344148

Julsez, B. (1971). *Foundations of Cycoplean Perception*. Chicago: University of Chicago Press.

Koenderink, J. J. (1987). "An internal representation of solid shape based on the topological properties of the apparent contour," in *Image Understanding 1985-1986*, eds. W. Richards and S. Ullman (Norwood, NJ: Ablex) 257–285.

Koenderink, J. J. (1990). *Solid Shape*. Cambridge, MA: MIT Press.

Koenderink, J. J., and van Doorn, A. J. (1975). Invariant properties of the motion parallax field due to the movement of rigid bodies relative to the observer. *Optica Acta* 22, 773–791. doi: 10.1080/713819112

Koenderink, J. J., and van Doorn, A. J. (1980). Photometric invariants related to solid shape. *Optica Acta* 27, 981–996. doi: 10.1080/713820338

Koenderink, J. J., and van Doorn, A. J. (1991). Affine structure from motion. *J. Opt. Soc. Am. A* 8, 377–385. doi: 10.1364/JOSAA.8.000377

Koenderink, J. J., and van Doorn, A. J. (1992a). Second-order optic flow. *J. Opt. Soc. Am. A* 9, 530–538. doi: 10.1364/JOSAA.9.000530

Koenderink, J. J., and van Doorn, A. J. (1992b). Generic neighborhood operators. *IEEE Trans. Patt. Analy. Mach. Intell.* 14, 597–605. doi: 10.1109/34.141551

Kowler, E. (2011). Eye movements: The past 25 years. *Vision Res.* 51, 1457–1483. doi: 10.1016/j.visres.2010.12.014

Lappin, J. S., and Bell, H. H. (1976). The detection of coherence in moving random-dot patterns. *Vision Res.* 16, 161–168. doi: 10.1016/0042-6989(76)90093-6

Lappin, J. S. and Bell, H. H. (2021). Form and function in information for visual perception. *I-Perception.* 12, 1–22. doi: 10.1177/20416695211053352

Lappin, J. S., and Craft, W. D. (2000). Foundations of spatial vision: From retinal images to perceived shapes. *Psychol. Rev.* 107, 6–38. doi: 10.1037/0033-295X.107.1.6

Lappin, J. S., Doner, J., and Kottas, B. L. (1980). Minimal conditions for the visual detection of structure and motion in three dimensions. *Science.* 209, 717–720. doi: 10.1126/science.7394534

Lappin, J. S., Donnelly, M. P., and Kojima, H. (2001). Coherence of early motion signals. *Vis. Res.* 41, 1631–1644. doi: 10.1016/S0042-6989(01)00035-9

Lappin, J. S., and Kottas, B. L. (1981). The perceptual coherence of moving visual patterns. *Acta Psychol.* 48, 163–174. doi: 10.1016/0001-6918(81)90058-5

Lappin, J. S., Norman, J. F., and Mowafy (1991). The detectability of geometric structure in rapidly changing optical patterns. *Perception* 20, 513–528. doi: 10.1068/p200513

Lappin, J. S., Nyquist, J. B., and Tadin, D. (2004a). Acquiring visual information from central and peripheral fields [Abstract]. *J. Vision* 5, 161. doi: 10.1167/5.8.161

Lappin, J. S., Nyquist, J. B., and Tadin, D. (2004b). Coordinating central and peripheral vision of stationary and moving patterns [Abstract]. *J. Vision* 4, 61. doi: 10.1167/4.11.61

Lappin, J. S., Tadin, D., Nyquist, J. B., and Corn, A. L. (2009). Spatial and temporal limits of motion perception across variations in speed, eccentricity, and low vision. *J. Vision.* 30, 1–14. doi: 10.1167/9.1.30

Lappin, J. S., Tadin, D., and Whittier, E. J. (2002). Visual coherence of moving and stationary image changes. *Vision Res.* 42, 1523–1534. doi: 10.1016/S0042-6989(02)00062-7

Legge, G. E., and Campbell, F. W. (1981). Displacement detection in human vision. *Vision Res.* 21, 205–213. doi: 10.1016/0042-6989(81)90114-0

Marr, D. (1982). *Vision.* San Francisco: W. H. Freeman.

Martinez-Conde, S., Macknik, S. L., and Hubel, D. (2004a). The role of fixational eye movements in visual perception. *Nat. Rev. Neurosci.* 5, 229–241. doi: 10.1038/nrn1348

Martinez-Conde, S., Macknik, S. L., Troncoso, X. G., and Hubel, D. (2004b). Microsaccades: a neurophysiological analysis. *Trend Neurosci.* 32, 463–474. doi: 10.1016/j.tins.2009.05.006

McKee, S. P., Welch, L., Taylor, D. G., and Bowne, S. F. (1990). Finding the common bond: Stereoacuity and the other hyperacuities. *Vision Res.* 30, 879–891. doi: 10.1016/0042-6989(90)90056-Q

Pack, C. C., Hunter, J. N., and Born, R. T. (2005). Contrast dependence of suppressive influences in cortical area MT of alert macaque. *J. Neurophysiol.* 93, 1809–1815. doi: 10.1152/jn.00629.2004

Perotti, V. J., Todd, J. T., Lappin, J. S., and Phillips, F. (1998). The perception of surface curvature from optical motion. *Percept. Psychophys.* 60, 377–388. doi: 10.3758/BF03206861

Ratliff, F. (1965). *Mach Bands: Quantitative Studies on Neural Networks in the Retina.* San Francisco: Holden-Day.

Reichardt, W. (1961). "Autocorrelation: A principle for the evaluation of sensory information by the central nervous system," in *Sensory communication,* ed. W.A. Rosenblith (New York: Wiley) 303–318.

Rodieck, R. W. (1998). *The First Steps in Seeing.* Sunderland, MA: Sinauer.

Rogers, B. J., and Graham, M. (1979). Motion parallax as an independent cue for depth perception. *Perception* 8, 125–134. doi: 10.1068/p080125

Rucci, M., Iovin, R., Poletti, M., and Santini, F. (2007). Miniature eye movements enhance fine spatial detail. *Nature* 447, 851–855. doi: 10.1038/nature05866

Rucci, M., and Poletti, M. (2015). Control and functions of fixational eye movements. *Ann. Rev. Vision Sci.* 1, 499–518. doi: 10.1146/annurev-vision-082114-035742

Strauss, S., Korympidou, M. M., Ran, Y., Franke, K., Schubert, T., Baden, T., et al. (2022). Center-surround interactions underlie bipolar cell motion sensitivity in mouse retina. *Nat. Communic.* 13, 1–18. doi: 10.1038/s41467-022-32762-7

Strogatz, S. H. (2003). *Sync: How Order Emerges From Chaos in the Universe, Nature and Daily Life.* New York: Hyperion.

Strogatz, S. H. (2018). *Nonlinear Dynamics and Chaos, second edition.* Boca Raton, FL: CRC Press. doi: 10.1201/9780429492563

Tadin, D. (2015). Suppressive mechanisms in visual motion processing: From perception to intelligence. *Vision Sci.* 115A, 58–70. doi: 10.1016/j.visres.2015.08.005

Tadin, D., and Lappin, J. S. (2005). Optimal size for perceiving motion decreases with contrast. *Vision Res.* 45, 2059–2064. doi: 10.1016/j.visres.2005.01.029

Tadin, D., Lappin, J. S., Gilroy, L. A., and Blake, R. (2003). Perceptual consequences of centre-surround antagonism in visual motion processing. *Nature.* 424, 312–315. doi: 10.1038/nature01800

Tadin, D., Park, W.-J., Dieter, K. C., Melnick, M. D., Lappin, J. S., and Blake, R. (2019). Spatial suppression promotes rapid figure-ground segmentation of moving objects. *Nat. Commun.* 10, 2732. doi: 10.1038/s41467-019-10653-8

Tadin, D., Silvanto, J., Pascual-Leone, A., and Batelli, L. (2011). Improved motion perception and impaired spatial suppression following disruption of cortical area MT/V5. *J. Neurosci.* 31, 1279–83. doi: 10.1523/JNEUROSCI.4121-10.2011

Todd, J. T., and Norman, J. F. (1991). The visual perception of smoothly curved surfaces from minimal apparent motion sequences. *Percept. Psychophys.* 50, 509–523. doi: 10.3758/BF03207535

Tyler, C. W., and Julesz, B. (1978). Binocular cross-correlation in time and space. *Vision Res.* 18, 101–105. doi: 10.1016/0042-6989(78)90083-4

Uttal, W. R. (1975). *An Autocorrelation Theory of Form Detection.* Hillsdale, NJ.

van de Grind, W. A., Koenderink, J. J., and van Doorn, A. J. (1992). Viewing distance invariance of movement detection. *Exper. Brain Res.* 91, 135–150. doi: 10.1007/BF00230022

van de Grind, W. A., van Doorn, A. J., and Koenderink, J. J. (1983). Detection of coherent movement in peripherally viewed random-dot patterns. *J. Optical Soc. Am.* 73, 1674–1683. doi: 10.1364/JOSA.73.001674

van Doorn, A. J., and Koenderink, J. J. (1982a). Temporal properties of the visual detectability of moving spatial white noise. *Exper. Brain Res.* 45, 179–188. doi: 10.1007/BF00235777

van Doorn, A. J., and Koenderink, J. J. (1982b). Spatial properties of the visual detectability of moving spatial white noise. *Exper. Brain Res.* 45, 189–195. doi: 10.1007/BF00235778

van Doorn, A. J., Koenderink, J. J., and van de Grind, W. A. (1985). Perception of movement and correlation in stroboscopically presented noise patterns. *Perception* 14, 209–224. doi: 10.1068/p140209

Westheimer, G. (1975). Visual acuity and hyperacuity. *Invest. Ophthalmol. Visual Sci.* 14, 570–572.

Westheimer, G. (1979). The spatial sense of the eye. Proctor Lecture. *Invest. Ophthalmol. Visual Sci.* 18, 893–912.

Westheimer, G., and McKee, S. P. (1977). Perception of temporal order in adjacent visual stimuli. *Vision Res.* 17, 887–892. doi: 10.1016/0042-6989(77)90062-1

Williams, D. W., and Sekuler, R. (1984). Coherent global motion percepts from stochastic local motions. *Vision Res.* 24, 55–62. doi: 10.1016/0042-6989(84)90144-5

frontiers | Frontiers in Computer Science

TYPE Review
PUBLISHED 21 June 2023
DOI 10.3389/fcomp.2023.1136987

OPEN ACCESS

EDITED BY
Mary Peterson,
University of Arizona, United States

REVIEWED BY
Jack Gallant,
University of California, Berkeley, United States
Naoki Kogo,
Radboud University, Netherlands
Matthew Self,
Netherlands Institute for Neuroscience (KNAW), Netherlands

*CORRESPONDENCE
Rüdiger von der Heydt
✉ rudiger8@gmail.com

RECEIVED 03 January 2023
ACCEPTED 30 May 2023
PUBLISHED 21 June 2023

CITATION
von der Heydt R (2023) Visual cortical processing—From image to object representation. *Front. Comput. Sci.* 5:1136987. doi: 10.3389/fcomp.2023.1136987

Visual cortical processing—From image to object representation

Rüdiger von der Heydt*

Department of Neuroscience and Krieger Mind/Brain Institute, Johns Hopkins University, Baltimore, MD, United States

Image understanding is often conceived as a hierarchical process with many levels, where complexity and invariance of object representation gradually increase with level in the hierarchy. In contrast, neurophysiological studies have shown that figure-ground organization and border ownership coding, which imply understanding of the object structure of an image, occur at levels as low as V1 and V2 of the visual cortex. This cannot be the result of back-projections from object recognition centers because border-ownership signals appear well-before shape selective responses emerge in inferotemporal cortex. Ultra-fast border-ownership signals have been found not only for simple figure displays, but also for complex natural scenes. In this paper I review neurophysiological evidence for the hypothesis that the brain uses dedicated grouping mechanisms early on to link elementary features to larger entities we might call "proto-objects", a process that is pre-attentive and does not rely on object recognition. The proto-object structures enable the system to individuate objects and provide permanence, to track moving objects and cope with the displacements caused by eye movements, and to select one object out of many and scrutinize the selected object. I sketch a novel experimental paradigm for identifying grouping circuits, describe a first application targeting area V4, which yielded negative results, and suggest targets for future applications of this paradigm.

KEYWORDS

visual cortex, figure ground organization, neural mechanism, object individuation, object permanence, selective attention, spiking synchrony, computational model

Introduction

We take it for granted that we see a world full of objects. But the images taken in by the eyes are just arrays of millions of pixels, and detecting objects from these arrays is a formidable task. It seems that the visual brain effortlessly provides us a representation of objects. Looking at Figure 1A, for example, we can easily answer questions like, what is the number of objects? how many corners has the green object? what is the color of the squares? which object is in the back? We can also compare two objects, or scrutinize a large complex object with multiple fixations. And when the display of Figure 1A is followed by the display of Figure 1B, we know that one object has moved from left to right. We have no doubt that it was one of the blue squares, although it is now neither blue nor a square. Complex natural images are certainly more difficult to process than the displays of Figure 1, but to understand vision, it seems to me, we should first understand how the visual brain enables us to make those assertions from such simple displays. What are the mechanisms that allow the brain to individuate objects from the stream of pixels, and how do they preserve their identity

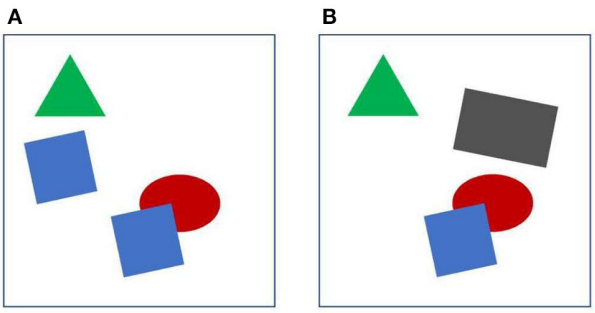


FIGURE 1
Object individuation. See text for further explanation. When **(A, B)** are displayed in sequence, one of the blue squares appears to move and at the same time change shape and color.

when the objects move? How do they achieve perceptual stability across eye movements, and how do they enable selective attention? This paper reviews studies that tries to answer those questions.

How does the visual cortex individuate objects? Since Hubel and Wiesel discovered the feature selectivity of simple, complex, and hypercomplex cells, early stages of visual cortex were thought to transform the pixel array into a representation of local image features like lines, edges, corners etc., which would then be assembled to larger entities that can be recognized as objects in inferior temporal cortex. This "hierarchical" scheme was questioned when a study showed that low-level cortical neurons that were supposed to signal lines and edges responded also to displays in which humans perceive illusory contours (von der Heydt et al., 1984). Contours are more than just edges and lines, they outline objects. At that time, illusory contours were commonly called "cognitive contours" because they appeared to be the result of a high-level, cognitive process, the system inferring a shape (like a triangle). Claiming that such contours are represented in a cortical area as low as V2 was to many a shock.

But the tide of vision sciences then had washed up the Fourier analyzer theory and neurophysiologists looked at the visual cortex as banks of spatial frequency filters. While this had the advantage of the convenient formalism of linear filtering, there were other indications (besides illusory contours) that cortical processing is highly non-linear from the beginning. In primary visual cortex it is not uncommon to find cells that respond to lines, but not to a grating of lines, and cells that respond vigorously to a sinusoidal grating of certain spatial frequency, but are totally unresponsive to the same grating when present as the 3rd harmonic component in a square wave grating (von der Heydt et al., 1992).

It took more than a decade until another perceptual phenomenon was found to have a correlate in visual cortex: figure-ground organization. Neurons in primary visual cortex respond to a texture in a "figure" region more strongly than to the same texture in a "ground" region (Lamme, 1995). Apparently, neurons at this low level already "know" what in the image is a figure, something that might be an object. But the tide of vision science then had surfaced another theory: coherent oscillations of neural firing were proposed to be the glue that holds the local features together as objects. Selective attention was thought to increase coherent oscillations which would lead to conscious perception. And the idea of the hierarchical scheme lives on in today's deep convolutional neural networks.

Border ownership coding

Neurophysiology led to another surprising discovery, neural selectivity for "border ownership" (Zhou et al., 2000). Figure 2 shows the basic finding. The responses of edge selective neurons, including the "simple" and "complex" types of Hubel and Wiesel, depend on how an edge is a feature of an object. The neuron illustrated responds strongly to the upper right edge of a square, and much less to the lower left edge. That is, the neuron responds to the identical local pattern differently, depending on whether it is an edge of an object to the bottom left of the receptive field, or an edge of an object to the top right. Indeed, for any location and orientation of receptive fields, there are two populations of neurons, those that "prefer" the object on one side of the receptive field, and those that prefer the object on the other side. Some respond also to lines, but many are strictly edge selective.

Zhou et al. termed this selectivity for "border ownership", adopting a term from the classic study by Nakayama et al. (1989) for the phenomenon that stereoscopic cues that change the way a border is perceptually assigned also affect object recognition: recognition of partly occluded objects is little impaired if the borders between occluded and occluding regions are stereoscopically assigned to the occluding regions (rendering them foreground objects), but is strongly impaired if these borders are stereoscopically assigned to the visible regions of the object.

The bottom of Figure 2A shows the time course of the neuron's mean firing rates. Because for each neuron with a border ownership preference one can find another neuron with the opposite preference, the two raster plots and the corresponding red and blue curves can be conceived as the simultaneous responses of a pair of neurons of opposite border ownership preferences. We also refer to the difference between the two as the "border ownership signal" (Figure 2B, dashed line, shading indicates SEM; from Zhang and von der Heydt, 2010). The border ownership signal is delayed by only about 15 ms relative to the mean response (thin line). These are responses from V2 neurons; border ownership signals of V1 have a similar time course. Note that Lamme's figure enhancement effect (where neurons respond to texture elements inside a figure) emerges later, about 50 ms after the response onset (Lamme, 1995).

The neuron of Figure 2 and the border ownership data to be reviewed below were recorded in rhesus macaques, but there is no doubt that the human visual cortex also represents contours by pairs of neurons of opposite border ownership preferences. A powerful paradigm for revealing selective neural coding is to demonstrate an adaptation aftereffect, which is based on the fact that cortical neurons exhibit short-term depression. Sure enough, it turned out that the classic tilt aftereffect is border-ownership selective. After adapting to a tilted edge that is owned by a figure on one side, a negative tilt aftereffect appears when the adapted location is tested with edges of figures on the same side, but not when tested with figures on the other side. And by alternating both, side-of-figure and tilt, during adaptation, one can produce two simultaneous tilt aftereffects in opposite directions at the same

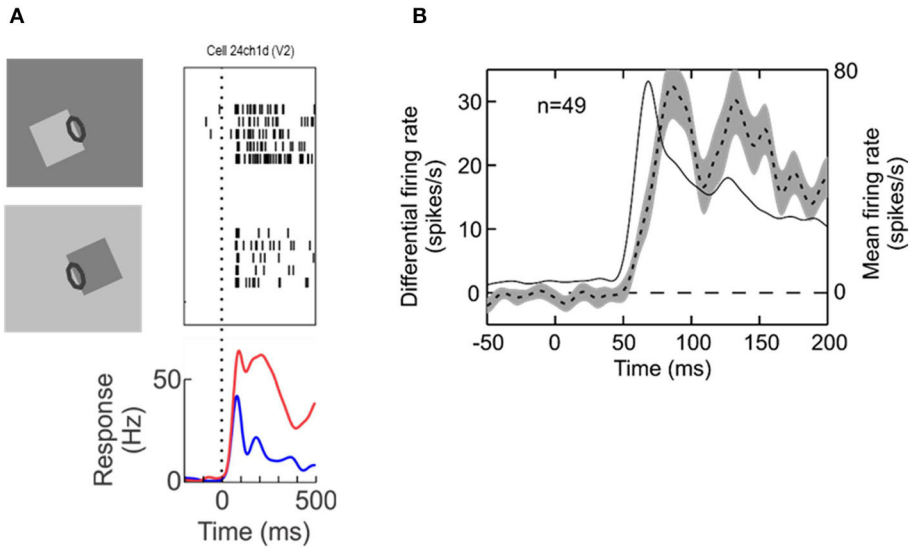


FIGURE 2
Border ownership selectivity. **(A)** The edge of a *square figure* is presented in the receptive field (*oval*) of a V2 neuron with the figure located either to the lower left or to the upper right. Note that the contrast was reversed between the two displays so as to compare locally identical stimulus conditions. Each trial started with a uniform field, and figure color and background color were both changed symmetrically at stimulus onset. The graphic depicts only tests with light-dark edges, but displays with reversed contrast were also tested, resulting in four basic conditions. *Raster plots* show the responses of the neuron, the curves show the time course of the mean firing rates. **(B)** *Dashed line*, time course of the difference between responses to preferred and non-preferred sides of figure—the "border ownership signal"—averaged across the neurons with significant effect of border ownership from one animal (*left ordinate*). *Shading* indicates standard error of the mean. *Thin solid line*, time course of responses (mean over the two figure locations, *right ordinate*).

location. Thus, there are two populations of neurons that can be adapted separately (von der Heydt et al., 2005).

Natural scenes

Are experiments with simple geometrical figures conclusive? The system may not need sophisticated algorithms to detect an isolated figure as in the displays of Figure 2. Other configurations that have been used in the early border ownership studies, like two overlapping figures, are also relatively simple compared to the complexity of natural scenes. Would neurons in V2 or V1 signal border ownership in natural scenes? Jonathan Williford tested neurons with large numbers of natural scenes (Williford and von der Heydt, 2016a). Using images from the Berkeley Segmentation Dataset (Martin et al., 2001) he selected many points on occluding contours for testing neurons (examples in Figure 3A). In the experiments, a fixation target for the monkey was embedded so that the selected points would be centered in the receptive field of the recorded neuron, and the image was rotated so that the contour matched the preferred orientation of the neuron. As in the standard border ownership test with squares, four conditions were tested: border ownership was controlled by rotating the image 180°, edge contrast was controlled by inverting the colors of the image so as to flip the colors between the regions adjacent to the contour. The data of this study are publicly available (Williford and von der Heydt, 2016b).

The first question was, can V2 neurons consistently signal border ownership under natural conditions? Each neuron was tested on many scene points (43 on average). The graph in Figure 3B shows the border ownership signals of an example neuron that was tested on 177 scene points. In seventy-nine percent of the cases the signals were consistent (plotted as positive in the graph). Consistency varied between neurons (Figure 3D). Out of 65, thirteen were over 80% consistent. In light of the hierarchical model of cortical processing, which is still widely accepted, the finding of consistent border ownership signaling in an area as low as V2 is highly surprising.

The cognitive hypothesis

The burning question is now, could border ownership modulation at this low level be the result of top-down projections from higher-level object recognition areas? Figure 4A shows a summary of the neuronal latencies (the time from stimulus onset to the beginning of responses) that have been reported for the various visual areas (after Bullier et al., 2001). One can see that neurons in object recognition areas in inferior temporal cortex (including IT, TE_x, TPO) respond relatively late. Of these, posterior IT (TPO) has the shortest latencies. To derive a prediction I use here the paper by Brincat and Connor (2006) who studied neuronal shape selectivity in the awake behaving conditions similar to those of the border ownership studies. Their study found that the response latencies in TPO depend on the type of responses within the area, with non-linear (shape selective) neurons having longer latencies than linear (unselective) neurons. The mean response for the shape selective group (green curve in their Figure 2B) reaches half-maximal strength at 130 ms. Thus, if border ownership selectivity in V2 depended on object recognition, the signal for natural scenes would

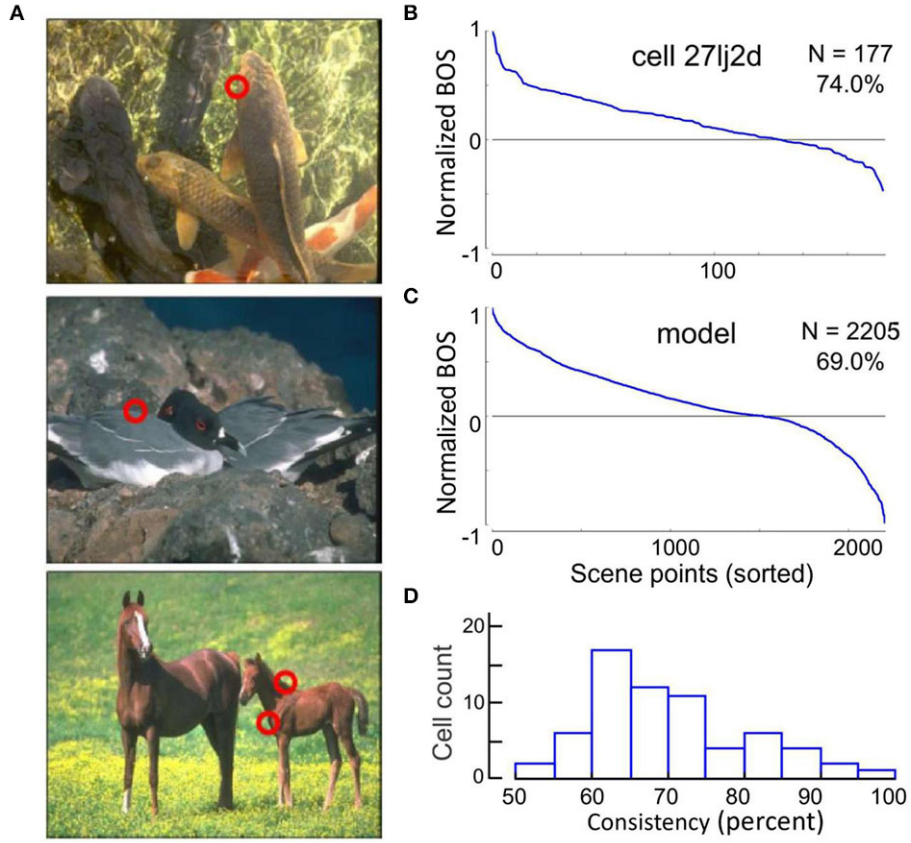


FIGURE 3
Neurons signal border ownership consistently across natural images. **(A)** Examples of images tested. *Red circles* show points where border ownership signals were measured ("scene points"). **(B)** Border ownership signals of an example V2 neuron normalized to the maximum and sorted. The signals were consistent for *74%* of the 177 scene points tested. **(C)** Performance of a computational model on the 2,205 scene points tested in the neurons. The model was consistent for *69%* of the points. **(D)** Distribution of the percentage of consistent signals of 65 V2 neurons with significant ($p < 0.01$) border ownership selectivity. Each neuron was tested with between 10 and 177 scene points (mean 43).

reach half-maximal strength only at 130 ms (or later, depending on delays added by the projection down to V2). Figure 4B "Prediction" shows how the earliest border ownership signals would then look like for natural scenes (red line) compared to the signals for displays of squares (dashed black line, half-max strength at 68 ms according to Zhou et al., 2000). What the experiment actually showed was that the border ownership signals for the two kinds of displays rise simultaneously (Figure 4B, *Data*) (Williford and von der Heydt, 2016a). We conclude that the cognitive explanation is untenable. The border ownership signals are faster than shape recognition in IT. This is the beauty of neurophysiology: it can easily rule out alternative hypotheses that would be difficult to discriminate with psychological or computational arguments.

What is the role of selective attention?

Attentive enhancement might be a plausible explanation for the figure-enhancement effect. When a figure pops up, it automatically attracts attention. But if a neuron responds more to a figure when it pops up here than when it pops up there, that difference cannot be the result of attention. The two displays in Figure 2 both contain a figure, the figure in the bottom display being flipped about the edge in the receptive field relative to the top display, and some neurons preferred one location, while others preferred the other location. It's a property of the neurons. Qiu et al. (2007) showed that border ownership and attentional modulation are separable aspects of neuronal function, and discovered an interesting correlation.

When the display contained several separate figures, and the monkey attended to one or another, border ownership modulation was found whether the figure at the receptive field was attended or ignored; there was only a slight difference in strength of modulation (Figure 5A).

And yet, attention does modulate the responses in displays in which objects partially occlude one another, and it interacts with border ownership in an interesting way. Figure 5B shows at the *top* the responses of an example neuron to the occluding contour. The two border ownership configurations are represented *left* and *right*, and side of attention in *top* and *bottom* rows. One can see that left was the preferred side of border ownership, and that the responses were enhanced when attention was on the left-hand object, compared to the right-hand object, for both border ownership conditions. Thus, attention on the neuron's preferred border ownership side enhanced the responses relative to attention on the non-preferred side, irrespective of the direction occlusion.

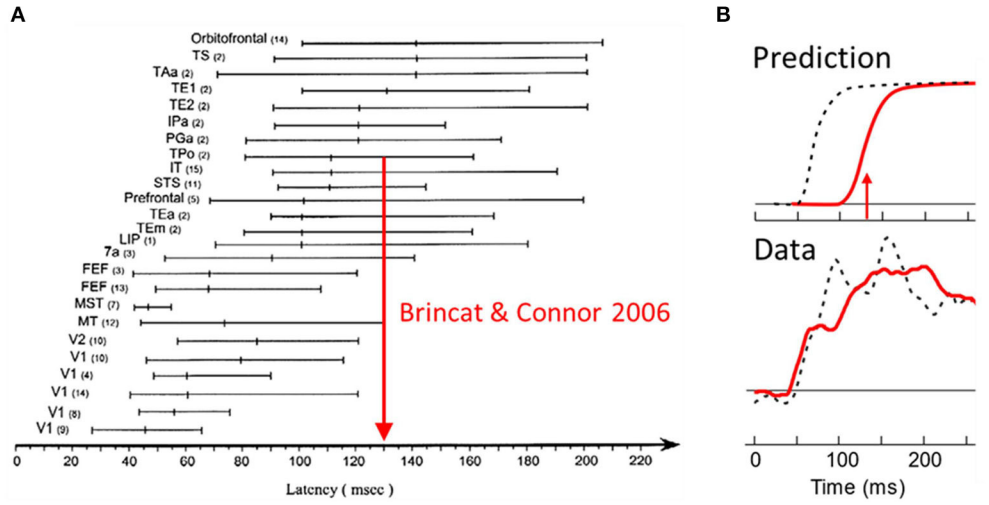


FIGURE 4
Selectivity of V2 neurons for border ownership in natural scenes cannot be the result of back-projections from object recognition centers in the inferotemporal cortex ("cognitive explanation") because it appears well-before shape selective responses emerge in inferotemporal cortex. **(A)** Summary of visual latency data across brain areas in monkey. Shape selectivity occurs first in posterior temporal cortex (*TPo*); *arrow* shows time of half-maximal strength of mean response of shape selective cells from a study in behaving monkey. **(B)** The border ownership signals for squares (*black dashed lines*) and natural scenes (*red solid lines*), as predicted, and as observed.

But attention did not override the border ownership signal. The results of Figure 5B, while showing the responses of one neuron to the two directions of border ownership, can be interpreted as the responses of two neurons with opposite border ownership preferences, which shows that, whether attention is on the left-hand object (*top* row) or on the right-hand object (*bottom* row), responses are stronger when the left-hand object owns the border.

Like in this neuron, the rule is that attentive enhancement is on the preferred side of border ownership, as shown by the *scatter plot* at the bottom of Figure 5B. The two factors were roughly additive, but there was a small but significant positive interaction. That is, attention enhanced responses more on the foreground object than on the background object.

The reader can experience the attention effect when looking at pictures in which border ownership is ambiguous. Figure 6 shows an artist's depiction of Napoleon's tomb on St. Helena. And not only his tomb, also his ghost, standing beside the tomb. To see him, direct your attention to the space between the trees!—The shape pops out because, when you first look at the picture, the neurons representing the borders between trees and sky are biased so that those assigning ownership to the tree regions prevail (smaller regions produce stronger border ownership signals than larger regions; the Gestalt Law of Proximity). But when their opponent neurons are enhanced by attention, ownership shifts to the sky region, and you can perceive its shape: the ghost.

The grouping cell hypothesis

I do not see the practical value of having the attention mechanism interfere with border ownership coding—besides the ability to see ghosts—but the linkage between attention effect and ownership preference helps in identifying the mechanism of border ownership selectivity. This linkage was a surprise because selective attention effects are usually phrased in terms of *regions* (left vs. right hemifield, figure vs. ground region) rather than *borders*.

How does a neuron of V1 or V2 know that the edge stimulating its receptive field is part of a figure? Could it be that border ownership selective neurons in V1/V2 are just Hubel and Wiesel's simple and complex cells that receive an additional modulating input from cells with large receptive fields that sense the presence of a big shape that might be an object? And that this modulating circuit is also used in top-down selective attention? That might explain why the attention effect is asymmetric about the receptive field, producing enhancement of responses when the attended object is on the preferred side of ownership.

The receptive fields of the neurons studied were near-foveal and typically about 0.5 deg in diameter, whereas the squares used to demonstrate border ownership selectivity measured 4 deg on a side or more. The neurons must be sensitive to the context far beyond the classical receptive field. Figure 7 illustrates an experiment in which the context influence was explored (Zhang and von der Heydt, 2010). The little gray specks left and right of the calibration mark show the classical receptive field of the neuron studied, and the vertical lines through the receptive fields depict the edges of the square stimuli, separately for the figure-left and figure-right conditions (the plot combines the results of two experiments, one with a 4° square, and one with a 7° square). To demonstrate the context effect, the figures were fragmented into eight pieces which were presented in random combinations, one combination per trial.

The top plot corresponds to the trials in which the various combinations of the contextual fragments were presented in addition to the edge fragment in the receptive field (the "center edge" for short). The bottom plot shows the trials in which the same contextual fragments were presented without the center edge.

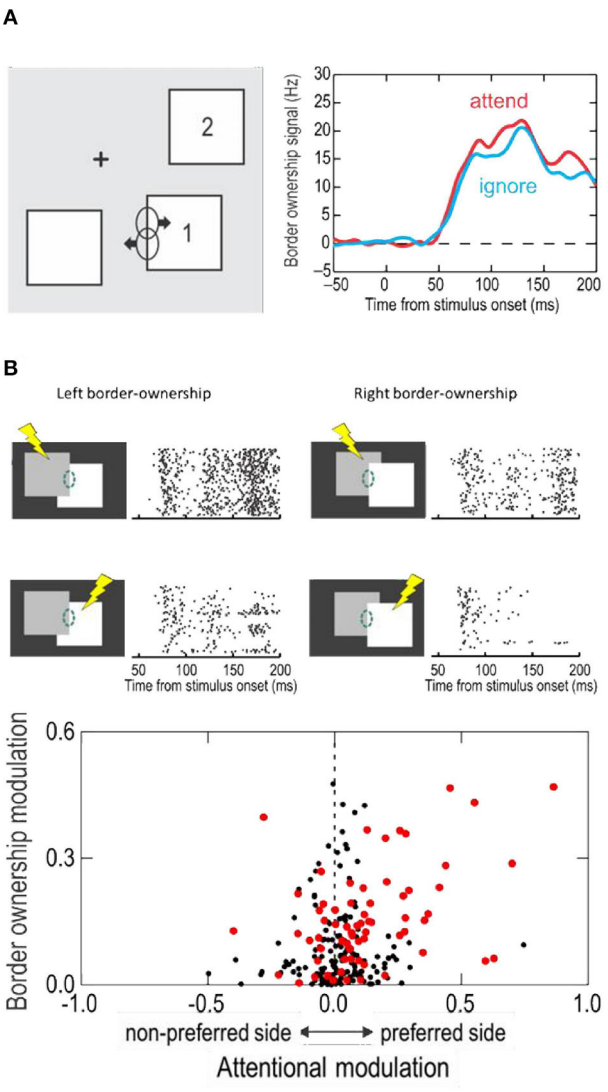


FIGURE 5
The influence of attention on border ownership signaling. **(A)** When objects are isolated, border ownership signals are similar whether the object at the receptive field is attended (*red line*) or ignored (*blue line*). **(B)** When objects overlap, side of attention and side of border ownership modulate the responses. *Raster plots* show responses of an example neuron. *Dashed oval*, receptive field. *Lightening symbols* indicate side of attention. *Scatter plot* shows the relation between attentional modulation and border ownership modulation across neurons of area V2. *Red dots* represent neurons that were influenced by both, attention and border ownership in form of significant main effects and/or significant interaction ($p < 0.01$).

FIGURE 6
Napoleon's tomb. Attending to a region flips border ownership to that region. See text.

The effect of each context fragment is indicated by color, *red* meaning enhancement of responses relative to the response to the center edge alone, *blue* meaning suppression. One can see that, for both figure sizes, the fragments to the left of the receptive field enhanced the center edge response, while the fragments to the right suppressed it. The bottom plot shows that the contextual fragments alone (without the center edge) did not evoke any responses.

The results from this kind of experiment show that, while neurons *respond* only to features within their small classical receptive fields, their responses can be *modulated* by the image context in a range that is much larger than the classical receptive field.

Nan R. Zhang also explored the context influence in the case of overlapping squares in which the border between the two squares (which perceptually belongs to the overlaying square) was placed in the receptive field. This situation is different in that there are figures on either side of the receptive field and mechanisms that simply detect the presence of a shape on one side would not work. The results showed that in this case the presence or absence of T-junctions, L-junctions (corners), and orthogonal edges, outside the receptive field modulated the responses to the center edge (von der Heydt and Zhang, 2018).

In area V4, where neurons are often selective for local contour features, Anitha Pasupathy and coworkers discovered that neurons that respond selectively to cusps are suppressed when the cusps are not object features, but accidental features produced by occlusion (Bushnell et al., 2011). Border ownership also affected the responses of shape selective neurons in infero-temporal cortex (Baylis and Driver, 2001).

The studies summarized so far led to the hypothesis that border ownership selectivity involves "grouping cells" that sum responses of feature neurons (including simple and complex types) and, via back projection, facilitate the responses of the same feature neurons, as sketched in Figure 8. Craft et al. (2007) designed a computational model in which grouping cells have fuzzy annular summation templates that are selective for oriented feature signals of roughly co-circular configuration. The summation of feature signals is linear, the feedback to the feature neurons is multiplicative. For example, the *blue G* cell in Figure 8 sums the responses of orientation selective neurons with receptive fields depicted in *blue*, and enhances their responses

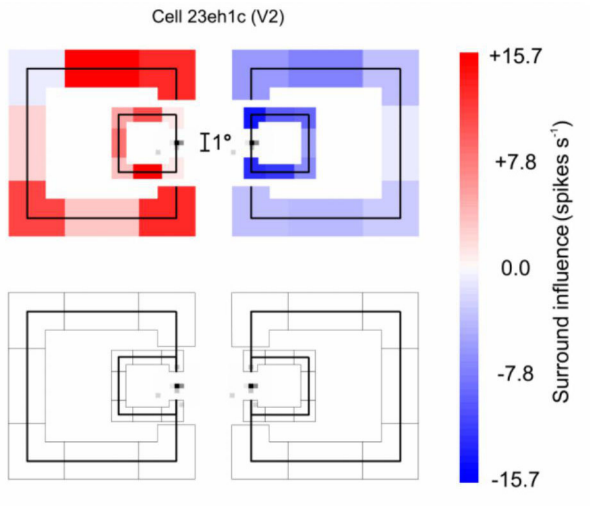


FIGURE 7
The extent of context integration in border ownership selectivity. This experiment probed locations where the responses of a border ownership selective neuron to a local stimulus could be modulated by contextual contours. First, the "classical receptive field" was mapped with a small probe stimulus. The map (*Gray specks*) is depicted twice to allow graphic representation of the two border ownership conditions. The contours of a square figure were fragmented into eight pieces (four edges and four corners) that were presented in random combinations, one combination at a time. One edge of the square was centered in the receptive field to stimulate the neuron, the other "contextual" fragments were presented outside the receptive field, as depicted in *red* and *blue* (they appeared either to the *left* or to the *right* of the receptive field, depending on the border ownership condition). The graph combines the results of two experiments, with 4° and 7° squares, respectively. The contextual fragments modulated the responses as indicated by the color of the fragments, *red* for facilitation, *blue* for suppression. One can see that, despite the tiny size of the classical receptive field, the contextual stimuli modulated the responses almost everywhere, those on the left producing facilitation and those on the right producing suppression. The *bottom graph* shows the result of control trials in which the edge in the receptive field was absent: none of the contextual fragments produced a response by itself.

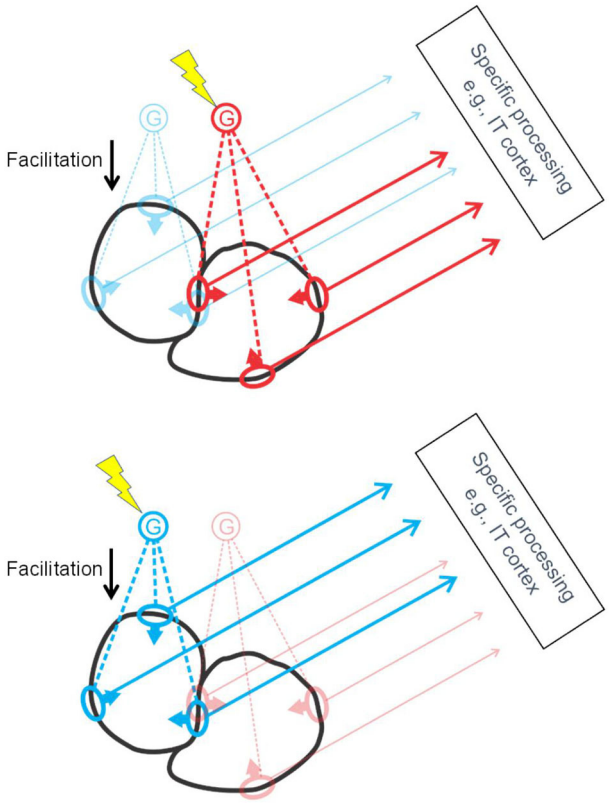


FIGURE 8
The grouping cell hypothesis. Schematic depiction of cortical activation by two objects (*black outlines*). The object contours stimulate receptive fields of edge selective feature neurons (*ovals*) which send their signals to specific processing centers at higher levels, and, via collaterals, also excite grouping cells *G*. Grouping cells project back to the same feature neurons, facilitating their responses. To select an object, top-down attention signals excite the corresponding G cell (*lightning symbol*) thereby enhancing the feature signals of the object for further processing. Grouping cells are thought to sum also signals that are not evoked by object features but by background structures cut off by occlusion, such as line terminations, L-junctions, and T-junctions. These inputs are not depicted here for clarity.

by feedback ("Facilitation"). This feedback makes those neurons border-ownership selective, as indicated by *arrows* on the receptive field symbols. Grouping cells also sum signals that do not correspond object features but indicate the layout of objects in depth, such as stereoscopic depth and accidental features produced by interposition, and the neurons providing these signals do not receive modulatory feedback. For example, T-junctions, and termination of lines at the contour, and orthogonal edges which contribute to border ownership (von der Heydt and Zhang, 2018).

Note that each piece of contour is represented by two groups of feature neurons for the two directions of border ownership, as illustrated in Figure 8 by the *red* and *blue* receptive field symbols in the center. Of the two objects depicted in *black*, the one to the left will activate the *blue G* cell, the one to the right, the *red G* cell. Selective attention, which consists in top-down activation of G cells (*yellow lightning shapes*), can enhance either the feature signals of the left-hand object (Figure 8, bottom) or those of the right-hand object (Figure 8, top). I have previously suggested that activation of a G cell (bottom-up or top-down) represents a "proto-object" (von der Heydt, 2015). This term had already been used in psychological studies, implying a preliminary object representation that may later be completed. The steep onset and early peak of border ownership signals do not indicate gradual completion but a one-shot process. But inspecting an object with multiple fixations seems to accumulate information about the details of an object in some central representation, which looks like gradual completion of an object representation. So, the emerging border ownership modulation and the enhanced feature signals might well be called a "proto-object". Where the completion occurs in the brain, and how, are questions that are worthwhile investigating.

The grouping cell hypothesis proposes that G cells come with summation templates of different sizes to accommodate the variety of objects. There must be a gamut of template sizes, and templates of each size must cover the visual field densely. The numbers of G cells required might raise concerns, but that number is actually

quite small, much smaller than the number of feature cells. This is because G cell templates only have low spatial resolution, the "resolution of attention" (Intriligator and Cavanagh, 2001), which is about 20 times lower than the feature resolution of the system. This means that, for the smallest template size, covering the visual field densely requires 400 times fewer G cells that feature cells. And the numbers of G cells with larger templates decreases in inverse proportion to square of size.

Besides the size of their summation template and the preference for co-circular signals, the hypothetical G cells are not particularly selective. The summation templates are fuzzy. Thus, round shapes, squares, and triangles, would all activate the G cells about as much as the potato shapes depicted in Figure 8. G cells are not "grandmother cells". Detecting a grandmother requires selectivity for *conjunctions*; not every person with white hair is a grandmother. By contrast, summing of features in G cells is *disjunctive*. Changing just one little T junction can flip foreground and background. In the case of partially overlapping squares, the border ownership signals for the occluding contour were found to grow with the number of indicative features, but saturate early: on average, one single feature (any T junction, L junction, or orthogonal edge) already produced half maximal signal strength (von der Heydt and Zhang, 2018). Selectivity was also found for border ownership defined by stereoscopic cues (Qiu and von der Heydt, 2005), motion parallax (von der Heydt et al., 2003), transparent overlay (Qiu and von der Heydt, 2007), and display history (O'Herron and von der Heydt, 2011). In the spirit of the grandmother cells terminology, G cells might be termed "TSA cells": "if you see something, say something."[1]

What characterizes an object are its feature signals. By targeting one G cell, the top-down attention mechanism can simultaneously enhance a large number of feature signals that characterize the exact shape, color, etc. of the target object. The G cells are not in the object processing stream, they serve only as handles to pick objects and allow attentive selection to route feature information of individual objects to higher processing centers, like those in inferior temporal cortex. For example, to read out the color of one of the squares in Figure 1, attention would boost the activity of a G cell according to location, while activating at the same time a color processing center downstream. From the feature neurons that are enhanced by the G cell, which include many color-coded edge selective neurons (Friedman et al., 2003), the color processor will compute the color of that square. Similarly, activating other processing areas will identify object shape and other object attributes.

As mentioned, every border between image regions activates pairs of border ownership selective neurons with opposite preferences. One such pair is depicted in Figure 8, the pair with receptive fields on the border between the two objects. This is to illustrate a specific prediction of the hypothesis, namely that attention to one side only facilitates the neuron that prefers that side of ownership. Thus, the grouping cell hypothesis predicts the correlation that was experimentally observed (Figure 5B). It predicts a hundred percent, whereas the actual correlation was lower, which is most likely due to the presence of basic spatial attention mechanisms in addition to the grouping cell mechanism. Attention may involve the grouping cell mechanism only in situations where simple spatial selection is not feasible, such as situations of partial occlusion, where the occluding contour should not be conflated with features of the background object.

Computational modeling shows the advantage of grouping cells in selective attention (Mihalas et al., 2011). Different from spatial attention models, the grouping cell model automatically localizes and "zooms in" on structures likely to be objects. The top-down attention signal only needs to enhance the G cell activity broadly in the region to be attended, and the network will direct the activity to potential objects in that region and focus activity on the size of G cell templates that fit each object best. The model replicates findings of perceptual studies showing that "objectness" guides and captures attention.

The above models (Craft et al., 2007; Mihalas et al., 2011) work on synthetic images of simple geometric shapes. A fully image computable model of the grouping mechanism was created by Hu et al. (2019a) and applied to natural images. The model produced contours as well as border ownership. Although it has no free parameters, Hu et al. found its performance to be overall comparable to state-of-the-art computer vision approaches that achieved their performance through extensive training on thousands of labeled images, fitting large numbers of free parameters.

The Hu et al. model has three layers of cells with retinotopic receptive fields, Simple cells (*S*), Border-ownership cells (*B*), and Grouping cells (*G*). Each *S* cell excites pairs of *B* cells for the two possible directions of border ownership. *B* cells thus inherit their receptive field selectivity from the *S* cells. *G* cells sum *B* cell responses according to fuzzy annular templates selectively for "co-circularity". The model works in an iterative manner. A given *G* cell sums the responses of one of the two *B* cells from each position and orientation, and facilitates the same *B* cells by modulatory feedback (see Figure 8) and suppresses the partner *B* cells by inhibitory feedback. This is motivated by neurophysiological results showing that image fragments placed outside the classical receptive field of a border ownership neuron can cause enhancement of the neuron's activity when placed on its preferred side, and suppression if placed on its non-preferred side (see Figure 7; the suppression is not depicted in Figure 8 for clarity). The model uses a scale pyramid of *G* cell template sizes, and pools information across different scales in a coarse-to-fine manner, with information from coarser scales first being upsampled to the resolution of the finer scale before being combined additively. A logistic function enforces competition between B cells such that their total activity was conserved.

Comparing with the neurophysiological data on the 2205 scene points tested in Williford and von der Heydt (2016a), Hu et al. found that their model achieved 69% consistent border ownership assignment, which was typical for V2 neurons (Figure 3). But the neurons varied, and many were actually more consistent. The neuron tested with the most images was 79% consistent across 177 scene points, and some were >90% consistent. This is no surprise because the Hu et al. model is simple. As Craft et al. (2007) observed, having grouping cells sum co-circular edge signals alone will not assign border ownership correctly for overlapping

1 Slogan of the Transport Security Agency.

figures, which the neurons do resolve. In Craft's model, grouping cell summation also included T-junction signals. As we saw above, the neurophysiology of border ownership coding suggests that grouping cells integrate a variety of different features as figure-ground indicators (von der Heydt and Zhang, 2018), and there seems to be a diversity of grouping cells, each using a subset of potential indicators. As a result, the consistency of border ownership coding across images varies from cell to cell (Figure 3). As Hu et al. (2019a) show, there was little similarity between two neurons when comparing their border ownership signals on a common set of scene points. Even highly consistent neurons are not entirely consistent with each other.

The computation of border ownership in natural scenes might be improved by having grouping cells include also local figure-ground indicators, similarly as the Craft et al. model included T-junction signals in dealing with simple geometrical figures.

Evidence for grouping cells

Do grouping cells exist? The observations of border ownership selectivity and attentive selection could also be explained by other hypotheses, for example, by propagating convexity signals along contours (Zhaoping, 2005), or by feedback projections in the cortical hierarchy from high-level areas with large receptive fields down to low levels with small receptive fields (Jehee et al., 2007), or simply by the magic of coherent oscillations. But there is one specific prediction of the grouping cell hypothesis: the top-down facilitation of feature neurons should lead to spiking synchrony, because all feature neurons that receive input from the same grouping cell (or cells) receive the identical spike trains. More specifically, synchrony should occur only between border ownership selective neurons when responding to the same object (*Bound* condition, Figure 9A); and only between pairs of neurons with "consistent" border ownership preferences (*red dashed lines* in Figure 9A), but not between "inconsistent" pairs (*gray dashed lines*). The hypothesis further predicts that synchrony will be found between neurons that are widely separated in cortex, because the grouping cells must be able to encompass the images of extended objects represented retinotopically in visual cortex.

Anne Martin tested these predictions, which was a difficult task. First, it required simultaneous stable recordings from two distant neurons, both of which had to be border ownership selective. Second, the objects had to be shaped according to the positions and orientations of the receptive fields of the two neurons encountered (sometimes it was impossible to construct a simple figure that would stimulate both neurons).

The main results are shown in Figure 9B (Martin and von der Heydt, 2015). The three different display- and attention conditions are depicted schematically at the top. While the subject fixated gaze on a fixation target (*black dot*), three figures were presented so that the two receptive fields (*red ovals*) were either stimulated by the same figure (*Bound*) or by different figures (*Unbound*). Additionally, attention was controlled (*asterisk*) by having the subject detect the moment of a subtle modification of shape that occurred predictably in one of the figures. Below, the frequency of spike coincidences is plotted as a function of lag time, after correcting for random coincidences (a cross-correlation function

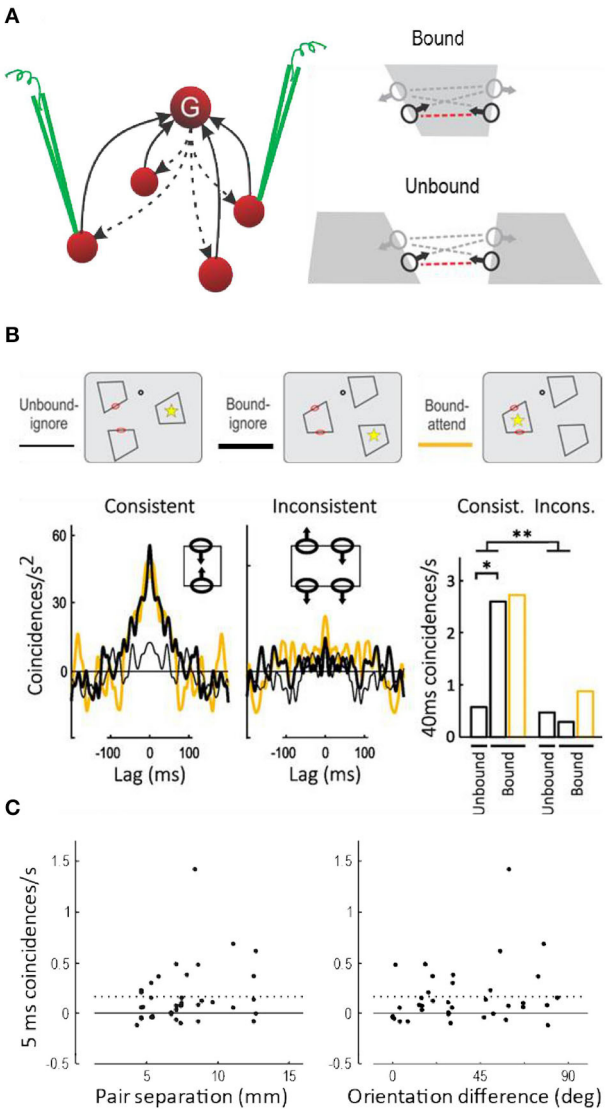


FIGURE 9
Spiking synchrony between border ownership selective neurons. **(A)** According to the hypothesis a single grouping cell *G* contacts many V1/V2 neurons via recurrent projections (*dashed arrows*). These neurons receive identical spike trains when the grouping cell fires, which should lead to spike synchronization. Because the grouping input produces border ownership preference, the hypothesis predicts synchrony between neurons whose border ownership preferences point toward the activating object (*consistent* pairs, indicated by *red dashed lines* between receptive field symbols), but only when they respond to the same object (*Bound*), and not when they respond to different objects (*Unbound*). **(B)** *Curves* show the covariograms between spike trains of pairs of neurons under the *experimental conditions* shown schematically above: *dot,* fixation point; *red ovals,* receptive fields; *yellow asterisk,* focus of attention (*ovals* and *asterisks* were not part of the display). Consistent pairs produced a sharp peak at zero (coincidence) when stimulated by the same object, whether the object was attended or ignored (*yellow and black heavy lines*), in contrast to stimulation with different objects which did not produce a peak (thin line). Inconsistent pairs produced rather flat covariograms. *Bar graphs* show the frequency of coincidences within 40 ms for the two kinds of pairs under the three experimental conditions. There was a significant difference between *Bound* and *Unbound* for consistent pairs, and a highly significant interaction between pair type and binding condition (This figure shows the data from the quartile of trials with the fastest
(Continued)

FIGURE 9 (Continued)
behavioral responses; results for all trials combined were similar except for an additional effect of attention, see Martin and von der Heydt, 2015). **(C)** Synchrony (frequency of 5-ms coincidences) as a function of the distance in cortex between the neurons of each pair (*left-hand plot*), and as a function of the difference between their preferred orientations (*right-hand plot*). *Dashed lines*, Mean. Neurons separated as widely as 13 mm fired synchronous spikes, as did neurons with different preferred orientations, indicating that individual grouping cells contact neurons representing distant features and features of different orientations.

called "covariogram"). There is a sharp peak at zero lag in the *Bound* condition, but not the *Unbound* condition; and for *Consistent* pairs, but not for *Inconsistent* pairs. The bar graphs to the right show the frequency of coincidences within 40 ms and the significance of the differences (the results were similar for 5-ms coincidences and the differences were also significant). This is exactly as predicted: in neurons that receive projections from a common grouping cell (i.e., neurons whose directions of border ownership preference are consistent) spiking synchrony increases when that grouping cell is activated (i.e., when both neurons are stimulated by a common object). I think the experiment cannot distinguish whether synchrony is due to single grouping cells or pools of such cells, but the sharp peak at zero lag of the covariograms in Figure 9B indicates coincidences of individual spikes.

Attention had little effect on synchrony (just as it produced little enhancement of responses, Figure 5A).

Spiking synchrony between neurons in primary visual cortex has generally been found to fall off rapidly with distance between neurons, reaching zero at 4 mm, which is approximately the maximum length of horizontal fibers in V1, and to be specific to neurons with like orientations (Smith and Kohn, 2008). The grouping hypothesis predicts the opposite: to be flexible, the grouping mechanisms must encompass neurons with widely separated receptive fields and a variety of orientation preferences. And indeed, in the above experiment, neurons separated by as much as 13 mm showed tight (5 ms) synchrony, and finding synchrony did not depend on similarity of preferred orientations (Figure 9C).

Is grouping behaviorally relevant? The task in the experiment of Figure 9 required detection of a small shape change produced by counterphase movements of the edges in the two receptive fields; the behavioral response depended on grouping these edges to one object. Thus, the hypothesis predicts that, if the strength of the grouping feedback fluctuates from trial to trial, stronger synchrony should be followed by a faster behavioral response. Anne Martin discovered that the response time correlated negatively with synchrony in consistent pairs in the "Bound" condition, whereas inconsistent pairs showed no such correlation. In the quartile of trials with the strongest synchrony the mean response time was 8 ms shorter than in the quartile with the weakest synchrony. Thus, the behavioral responses were fastest when neural grouping was strongest, as predicted.

One question we glanced over above is, how can *modulatory* common input produce synchrony? Spiking synchrony is generally observed when two neurons are *activated* by a common spike train, but, according to the theory, grouping cell feedback to feature neurons does not activate, but only *modulates* existing activity (see example in Figure 7 showing that context features alone do not activate). Nobuhiko Wagatsuma and Ernst Niebur explored synchrony between pairs of feature neurons with a spiking model. They modeled the afferent inputs by independent spike trains activating AMPA receptors, and the modulatory grouping cell input by a common spike train activating NMDA receptors (using a standard computational model for generic NMDA receptors). Surprisingly, this model produced synchrony, and even the exact shape of the experimental covariograms and the observed synchrony at millisecond precision (Wagatsuma et al., 2016).

As we have seen, experiments and modeling confirm a critical prediction of the grouping cell theory: that pairs of border ownership selective cells with consistent direction preferences, when activated by a common object, exhibit spike train synchrony with a cross-correlation function whose shape is characteristic for common modulatory input. Next, we will consider another critical prediction of the theory, persistence.

Persistence

It has been argued that vision—in contrast to audition—does not need short-term memory because the visual information is continuously available so that attention can always pick what is needed. But I argue that vision needs a short-term memory too. What would be the use of grouping features to objects if that would all be lost in a blink?

O'Herron and von der Heydt (2009) devised experiments to test if border ownership signals persist. The idea was to present an edge in the receptive field that is owned by a figure on one side, as in the standard test of Figure 2, and then, keeping the edge in the receptive field, switch to a display in which ownership of the edge is ambiguous. This simple paradigm has produced amazing results. Figure 10, top, shows the sequence schematically for ownership-left (the corresponding displays for ownership-right were also tested to measure the border ownership signal). Below, the *red curve* shows the average time course of the signal. It rises steeply and stays high during the figure phase, as in the standard test, but in the ambiguous phase it declines only slowly. For comparison, when the figure was flipped to the other side keeping the edge contrast (Figure 10, 2nd row of insets from top), the signal changed quickly to negative values (*blue curve*). The difference between the time constants was 20-fold. Thus, border ownership signals persist for a second or more. This experiment also shows that the persistence is not due to inherent persistence of responses in the recorded neurons, because in the "flip" condition their responses change rapidly.

The paradigm of Figure 10 is somewhat artificial in that it does not have a simple interpretation in terms of objects with natural continuity. In the top display sequence, the initially presented object disappears and a bipartite field appears, and in the sequence below, the initial object disappears, and a different object appears on the opposite side. To study persistence of border ownership signals in a more natural situation, Philip O'Herron designed an ingenious display sequence in which objects maintain continuity.

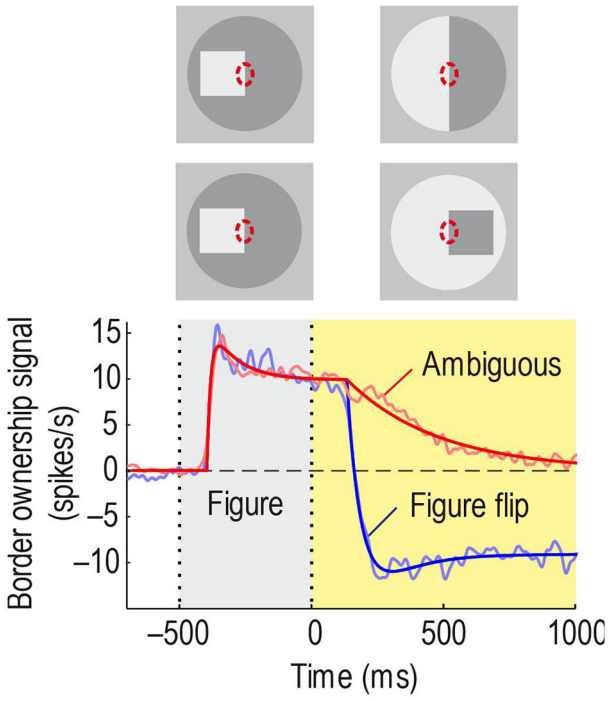


FIGURE 10
Persistence of border ownership signals. The edge of a square was presented in the receptive field for 0.5 s and then switched to an ambiguous edge (in other trials, side of figure and contrast were reversed, like in the standard test). In the other condition, the square was flipped to the other side. *Red line* shows the time course of the border ownership signal when the edge is made ambiguous: The signal decays slowly. *Blue line*, time course of the signal when the square was flipped to the other side: The signal reverses quickly.

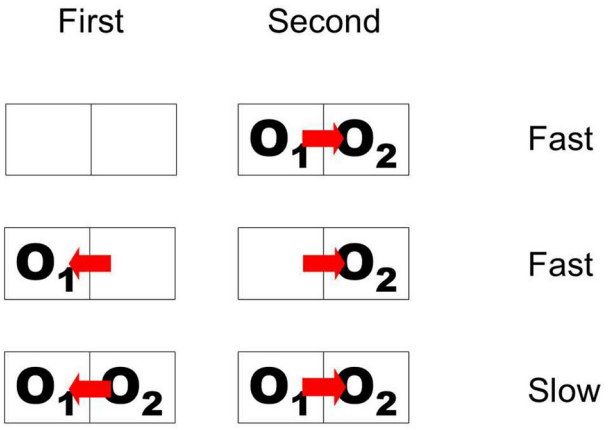


FIGURE 11
Schematic of display sequences producing fast and slow changes of border ownership signals. First and Second are two phases of stimulation. *Outlines* represent object locations; *O1* and *O2*, objects; *red arrow*, direction of border ownership. When overlapping objects appear suddenly, or when the object on one side disappears and another object appears on the opposite side, the signal changes fast. But when two objects are continuously present and ownership of the common border is reversed, the signal follows only slowly.

He presented two partially overlapping figures and recorded responses to the common border (the occluding contour) when the occlusion cues reversed while the two figures were continuously displayed. The result was that the initial border assignment persisted for 2 s or more before reversing sign (O'Herron and von der Heydt, 2011). Control conditions showed that, when the final configuration of overlapping figures was presented without history, the signal assumed the final value quickly; and when a single figure was presented on one side and was then replaced by a figure on the opposite side, as in Figure 10 *Figure flip*, the signal also reversed quickly. These results are summarized schematically in Figure 11, where pairs of adjacent *frames* represent two object locations, *O1* and *O2* denote two objects, and the *red arrow* indicates direction of border ownership. Abrupt-onset and object-flip result in fast signal changes, whereas reversal of occlusion cues in the presence of both objects results in retarded reversal of the signal. It seems that object continuity includes continuity of depth relations. More generally, we hypothesize that the system represents location in space as an object attribute which has continuity unless there is an abrupt image event like onset or offset.

O'Herron also showed that the persistent ownership signals "remap" across saccades, a result that will be reviewed below. The persistence of border ownership signals is another example of the power of neurophysiology in providing clear answers to questions that are difficult to answer with psychological methods.

How is it possible that neural signals rise fast and decay slowly? Neurons in low-level visual areas must be able respond fast to the afferent signals from the retina which change swiftly with new information arriving after a fraction of a second. The memory-like behavior shown in Figure 10 is a puzzle for neural network theory. Traditional positive feedback models show attractor dynamics, with transient perturbations resulting in a quasi-permanent change of system state, whereas the responses of Figure 10 return to the original state after a transient. This is a question of very general interest because short-term memory underlies many kinds of behavior. Grant Gillary discovered that short-term depression, which is ubiquitous among cortical neurons, can create short-term persistence in derivative feedback circuits. If short-term depression acts differentially on positive and negative feedback projections between two coupled neurons, they can change their time constant dynamically, allowing for fast onset and slow decay (Gillary et al., 2017).

The blessing and the curse of eye movements

We see by moving our eyes. The eyes fixate, producing stable images for a moment, and then move rapidly to fixate another part of the scene. Each time, the images are displaced in the eyes. Humans as well as monkeys move fixation continually about 3–4 times per second. The reason why primates do this is obviously to be able to scrutinize different parts of a scene with the high-resolution center region of the retina and its corresponding processing apparatus in the brain. The system then synthesizes information from multiple fixations to represent

complex objects and combines object representations to scene representations. From the computational standpoint this seems horribly complicated. At least I don't know of any technical system that would bounce a camera around four times per second. How our brain deals with this confusing input is a puzzle. One question is, why don't we perceive and are not disturbed by the frequent movements of the retinal image, but rather perceive a stable world. But the subjective stability is a minor issue compared to the question of how the system integrates object information across the eye movements.

To explain cross-saccadic integration, van Holst proposed the reafference principle (von Holst and Mittelstaedt, 1950). When the brain creates a signal that commands the eyes to move, he thought, it also produces an associated signal that tells the visual system about the impending eye movement and informs it about the direction and size of the image movement to expect. He called the change of retinal signals caused by the eye movement the "afference", and the associated brain signal to the visual system the "reafference". To create continuity the brain would have to correct the afference by the reafference, that is, to shift the image representation so as to cancel the image movement and thus achieve a stable internal representation (Figure 12A). The problem with this theory is that a shifter circuit that could remap the image representations would have to be huge. V1 and V2 each consist of over 100 million neurons and there is no other structure in the visual brain that could hold so much information.

An alternative solution would be to work with image representations that move with every eye movement, and remap the object structure accordingly (Figure 12B). Instead of requiring a stabilized image representation, object-based attention would then only need object pointers that are updated with every eye movement. Zhu et al. (2020) conjectured that top-down attention signals activate object pointer cells whose signals are fed via a shifter circuit to grouping cells. This scheme would reduce the stabilization task from remapping millions of image signals to remapping a few object pointer signals. Assuming the system can maintain a number of object pointers, top-down attention could select to which object to attend, and the remapping would preserve its identity and enable the attention mechanism to keep focused on it, that is, keep enhancing the feature signals of that object across eye movements, or deliberately choose to focus on another object.

Evidence for remapping of border ownership

The hypothesis of object pointer remapping implies that the activation of grouping cells is being remapped to a new location with each eye movement. When an object appears, a grouping cell responds and will activate an object pointer. This activity persists and, by feedback, reinforces the activity of the grouping cell. When the eyes then make a saccade that moves the image of that object to the receptive field of another grouping cell, the shifter circuit will reroute the object pointer accordingly and its activity will flow down to the new grouping cell. Thus, the grouping cell in the new location will become active immediately. The result will be that border ownership is remapped, that is, the feature neurons that respond to the object in the new location will be biased immediately, without the need for new context processing.

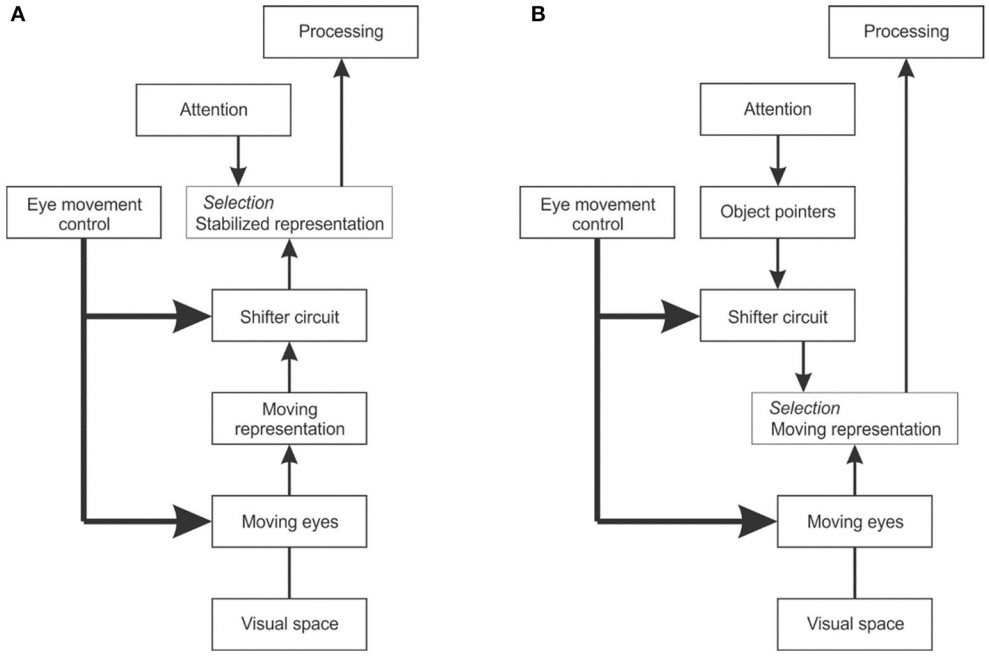


FIGURE 12
The problem of dealing with image displacements caused by eye movements. **(A)** Reafference principle. When the eye movement control center sends a command to the eye muscles, it also sends a copy to the visual system indicating the impending eye movement. A shifter circuit shifts the visual representation accordingly to compensate for the image displacement caused by the eye movement. **(B)** Object pointer hypothesis. The visual system works with the moving image representation, but uses object pointers that are updated at each eye movement by a shifter circuit.

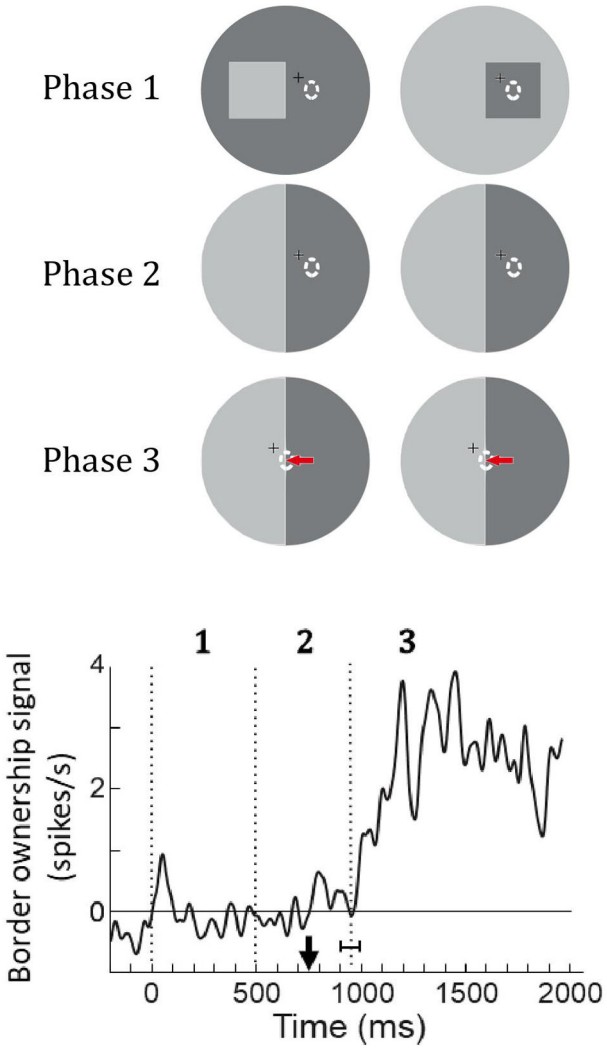


FIGURE 13
Remapping of border ownership signals across saccades. Like in the memory experiment of Figure 10, a figure was presented (*Phase 1*) and then replaced by an ambiguous edge (*Phase 2*), but in this experiment, the neuron was not stimulated because the edges were outside its receptive field (*dashed oval*). Finally, a saccade was elicited that brought the receptive field onto the ambiguous edge (*red arrow, Phase 3*). Curve shows the mean border ownership signal (47 neurons). *Black arrow* on time axis, movement of fixation point, *vertical dotted line* with *horizontal bracket*, mean time of saccade with standard deviation. The border-ownership signal is close to zero during *phases 1* and *2* because the neurons are not stimulated. But shortly after the saccade, when the edge stimulates the neurons, a signal emerges although border ownership is ambiguous. The neurons signal how the edge was owned before the saccade.

O'Herron and von der Heydt (2013) tested this prediction as illustrated in Figure 13. Recording from a feature neuron they presented a figure so that its edges were outside the receptive field (*Phase 1*) and then replaced the figure with an ambiguous edge that coincided with one of the figure edges (*Phase 2*). After a while, the fixation point was moved, inducing the monkey to make a saccade that brought the receptive field onto the edge (*Phase 3*). The prediction was that the neuron's responses will reflect the previous ownership despite the absence of a figure. The graph at the bottom shows the population border ownership signal. There are no responses in *Phases 1* and *2*, as expected, because the receptive fields are in a blank region. During *Phase 2*, the fixation point moves (*black arrow* on time axis) eliciting a saccade that brings the receptive fields onto the edge. The neurons respond, and a border ownership signal emerges as predicted. This is about half a second after the figure was removed; border ownership is produced from memory.

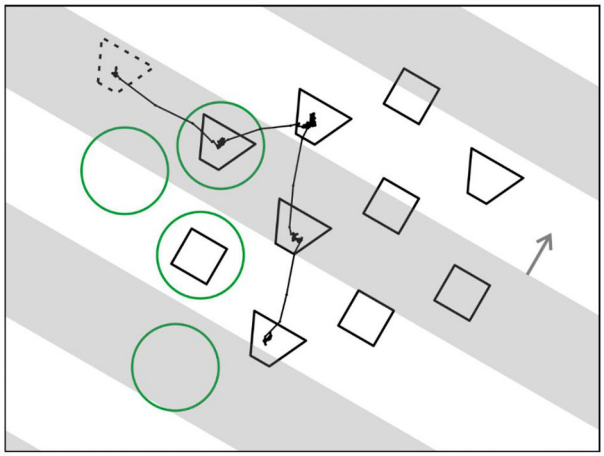


FIGURE 14
Searching for grouping cells—a novel paradigm. An array of *test figures* is presented, and monkey is instructed to scan the array in search for a figure that, when fixated, leads to reward. While monkey is scanning, opaque strips drift across the array (*gray bars*, depicted transparent for illustration only). For control of attention two shapes of figures are displayed, and a cue figure presented at the beginning of each trial (*dotted outline*) tells monkey which shape to look for. *Line trace*, eye movement trace of an example trial. *Green circles*, position of an example receptive field for each fixation.

Searching for grouping cells and object pointers

The results described so far are all based on variants of the border ownership paradigm and on recordings from V1, V2, and V4, and together they constitute strong evidence for the grouping cell theory. But the one crucial prediction of the theory, the existence of grouping cells has not yet been confirmed. Grouping cells might live in another brain region. In fact, finding persistence of the border ownership signal in areas like V1 and V2, where neuronal responses rise and fall fast, makes it seem unlikely to find grouping cells there.

Identifying grouping cells, to my knowledge, has only been attempted in one candidate area, V4, an area where some neurons have larger, fuzzy receptive fields and that has strong back projections to V2 and V1. Also, V4 is connected to both, the What and the Where pathways (Ungerleider and Mishkin, 1982; Ungerleider et al., 2008), and the function of grouping cells is just to pull out what is where.

Searching for grouping cells needs a different paradigm. The distinctive feature to look for is obviously the persistence of

responses when an activating object disappears from view, as in O'Herron's demonstration of persistence of border ownership where an edge is substituted for a square. But we do not expect grouping cells to respond to edges; rather, they should respond best when an object is centered on their summation template.

Alex Zhang and Shude Zhu developed a new paradigm motivated by the phenomenon that objects persist perceptually when they are transiently occluded, a phenomenon called "object permanence". When an object is occluded by another object passing in front of it and then reappears, we perceive it as the same object. We would be surprised if it had vanished, or if there were now two objects instead of one. The visual system holds the representations for a certain time even when the objects are invisible.

Figure 14 illustrates the new paradigm. The stratagem was to present an array of objects for visual search and, while the observing subject is scanning the array, transiently occlude some of the objects.

To control top-down attention, the objects were of two different shapes and, before the array appeared, a cue object was displayed (*dashed outline*) that specified which shape to look for. The cue object disappeared when the array came on.

Figure 14 also shows the example of an eye movement trace of a trial in which a trapezoidal shape was cued. The monkey made four fixations, and four *green circles* indicate where the receptive field of an example neuron would be in each case (the circles are only for illustration, they were not part of the display). In fact, the array was constructed for each neuron being recorded so that, when one of the objects was fixated, another object would fall on the neuron's receptive field in most trials, and in other trials, a blank region. In the example, two fixations brought objects into the receptive field, one a trapezoid, and the other a square, while in two other fixations the receptive field landed on a blank region.

Occlusion was added by having a series of opaque *gray strips* drift across the array that occluded half of it at any time (the *strips* are depicted as transparent in the Figure just for illustration; in fact, display items that we call "occluded" were physically absent). Surprisingly, the subjects had no difficulty in dealing with that complication. Once they mastered the task without occlusions, they rapidly adjusted to the occlusions in just one session. This of course confirms the power of perceptual permanence.

In the new paradigm neurons respond to static objects brought into their receptive fields by eye movements, much like in natural viewing, which is fundamentally different from the traditional neurophysiological paradigms in which neurons respond to objects that are being switched on and off. A technical complication here is that "stimulus onset" is not controlled by the experimenter, but by the subject's eye movements, which means that the neural responses are timed by onset and offset of fixation. Thus, the phases of visibility and occlusion of individual objects, which are programmed by the experimenter, need to be related to the recorded eye movements. But this complication is greatly outweighed by the opportunity to study neuronal activity under quasi natural viewing conditions which makes this an enormously powerful paradigm.

Studying V4 neurons with this paradigm Zhu et al. (2020) indeed found a "response" to the invisible objects in the mean firing rate, corresponding to the predicted top-down activation of grouping cells (their Figure 10 which shows the averaged responses of 87 V4 neurons). But the authors rejected this result as evidence for grouping cells in V4, suggesting an alternative explanation for the "responses" to invisible objects, because of another result that was not consistent with the predictions: While top-down attention and saccade planning clearly produced response enhancement for visible objects, they did not so for occluded objects (Figure 15).

Neurophysiology can be hard to understand if one just looks at what the various individual neurons do; only a theory can relate the neural signals to visual experience or the performance of a vision algorithm. Figure 15A illustrates the prediction of the theory when fixation is on one object (*square* marked by *yellow asterisk*) and a saccade is planned to another object (*dashed square*) that is momentarily occluded by a larger object (*blue outline rectangle*). According to the theory, there are three layers of cells, the feature cells with receptive fields in retinal space (*ovals on gray bars*), a grouping cell layer *G* with fixed connections to the feature cells, and a number of object pointer cells *OP* that are connected with grouping cells through the shifter circuit *SH*.

The top panel of Figure 15A illustrates the fixation before the saccade: top-down attention enhances the *OP* cell that is momentarily connected to grouping cell *G3*. Grouping cell *G1* (assumed to be the recorded cell) is not active because the object in its receptive field is occluded.

When the saccade to the other object is planned, as shown in the middle panel, top-down attention moves to the *OP* cell that is momentarily connected with *G1*, and the *OP* activity flows down to *G1* (*red arrow*). This is the predicted activity that will be recorded despite absence of afference from the retina.

And when the saccade is executed, as shown in the bottom panel, *SH* reroutes the connections to *G3* and *G5* as indicated by *yellow arrows*. Thus, while the left-hand object activates other feature cells after the saccade, it is again connected to the left-hand *OP* cell.

Figure 15B shows the time course of the mean firing rates at the end of a fixation period, that is, at the moment when the brain initiates a new saccade. The curves represent the activity from before the saccade until 50 ms after the saccade. Because 50 ms is the latency of visual responses in V4, visual information from the next fixation did not influence this activity.

The top three curves show the responses to visible objects, *red line* for responses when the object in the receptive field was the goal of the next saccade, and *brown lines* when another object was the goal; *solid lines*, when the object was a target, and *dashed line* when it was a distracter. These curves show that the responses were enhanced by attention (*solid brown* vs. *dashed brown*) and further enhanced when the attended object was the goal of the next saccade (*solid red* vs. *solid brown*). But planning a saccade to an occluded object did not produce the activity predicted by the red arrow in Figure 15A (*blue vs. cyan curves*) and occluded targets were not represented by enhanced activity compared to occluded distracters (*solid cyan* vs. *dashed cyan*). This means that the recorded neurons were activated by visual afference, but not by top-down activity from object pointer cells.

To conclude this section, previous experiments had shown that border ownership signals in neurons of V1/V2 persist after the

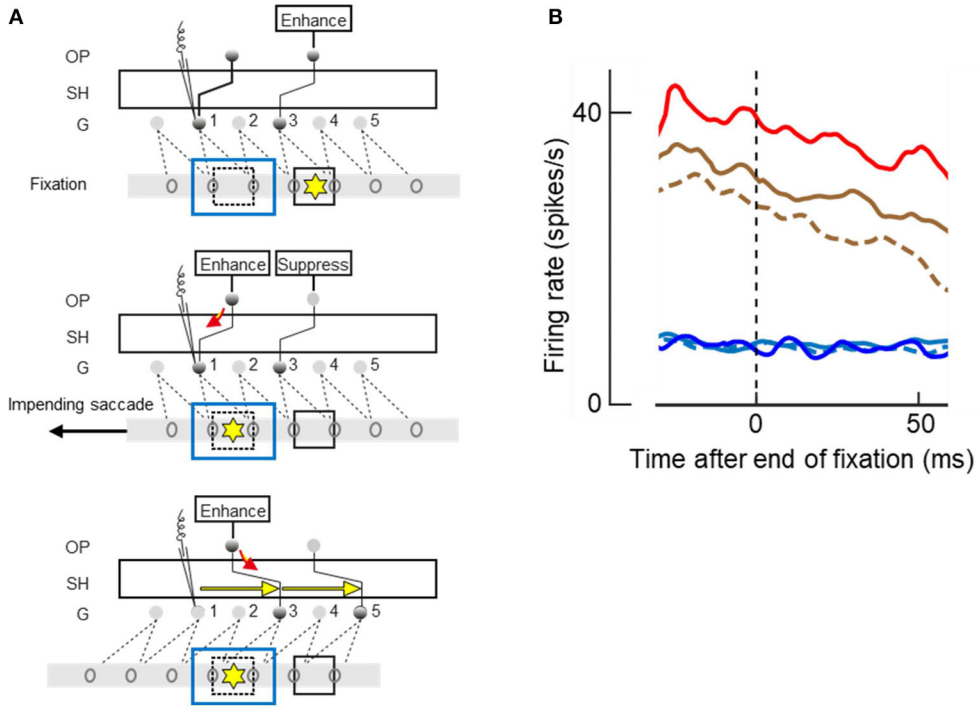


FIGURE 15
Attending and saccading to invisible objects. **(A)** Theory of object pointers. *Gray bars* with *ovals* represent receptive fields of feature cells in retinal space. *G*, grouping cell layer, *SH* shifter circuit, *OP* object pointer cells. Presentation of an object excites a number of feature cells and a grouping cell, and, through *SH*, an object pointer cell. *OP* cells sustain activity once excited. *Top panel*. Two objects (*squares*) have activated two *OP*s before the left-hand object was occluded by another object (*blue outline*). Attention on right-hand object (*yellow asterisk*), which is currently fixated, is enhancing corresponding *OP*. *Middle panel*. Planning of a saccade to left-hand object re-allocates attention, enhancing left-hand *OP* whose activity flows down to *G1* (*red arrow*), and a signal is recorded (*microelectrode symbol*) even though the object is no longer visible. *Bottom panel*. The saccade has moved the receptive fields, and SH has compensated for the movement by re-routing the connections to *G3* and *G5*, as indicated by *yellow arrows*, thus keeping left-hand *OP* connected to the feature cells of left-hand object. **(B)** The mean time course of activity recorded from 87 V4 neurons at the end of a fixation period; zero on abscissa marks time of saccade. Note that new visual input does not affect responses until ~50 ms, the latency of V4. *Red* and *brown* traces, responses to visible objects; *blue* and *cyan* traces, responses to occluded objects; *solid lines*, when attended; *dashed lines*, when ignored; *red* and *blue*, when goal of next saccade. Responses were enhanced by attention (*solid vs. dashed*), and further enhanced when object was goal of planned saccade (*red vs. brown*), but only for visible objects. Had the recordings been from a *G* cell, enhancements would also be found for occluded objects.

object that produced these signals has been removed (Figure 10), and that they even persist across a half second of display of a blank field that completely silences the activity of these neurons (O'Herron and von der Heydt, 2009, their Figure 7). These findings suggests that border-ownership selective neurons must be modulated by an external signal, by activity that we do not see in V1/V2. And the results of the new experiment, summarized in Figure 15B, show that this signal does not come from V4.

Plausibility of models

Since figure-ground organization was discovered by the Gestalt psychologists it has stimulated theories about the underlying brain activity unlike few other phenomena in perception, and the interest in modeling it has grown since neurophysiologists discovered neural activity related to illusory contours (von der Heydt et al., 1984), figure ground segregation (Lamme, 1995), object-based attention (Roelfsema et al., 1998), and border ownership (Zhou et al., 2000).

Among the various models of perceptual organization that have been proposed (Grossberg and Mingolla, 1985; Zhaoping, 2005; Jehee et al., 2007; Kogo et al., 2010; Jeurissen et al., 2016), the grouping cell model discussed here is distinct in that it makes the highly specific prediction that pairs of border-ownership selective neurons with consistent side-of-figure preferences, when stimulated by a common object, show spiking synchrony. And experiments have shown exactly this. Other neural models do not predict synchrony because the neurons representing the distributed features of an object are not supposed to receive input from common spike trains. Only the models by Jehee et al. (2007) and Jeurissen et al. (2016) propose neurons with receptive fields large enough to encompass objects. However, the coarse-to-fine processing in their model is relayed through a cascade of neurons down through the hierarchy of visual areas from TEO to V1, and the relays do not preserve spike timing.

Models that rely on lateral signal propagation (Grossberg and Mingolla, 1985; Zhaoping, 2005; Kogo et al., 2010) are not physiologically plausible because the conduction velocity of horizontal fibers in cortex is too slow. Based on published conduction velocity data, Craft et al. (2007) estimated that lateral

propagation would delay the border ownership signal for the 8 deg square by at least 70 ms relative to the edge responses, in addition to processing delays, whereas only 30 ms has been found. Sugihara et al. (2011) calculated the latencies of border ownership signals for two conditions in which the relevant context information was located at different distances from the receptive field and compared the latency difference with the difference predicted from horizontal signal propagation. The prediction was based on the increase in cortical distance computed from mapping of the actual test stimuli onto the cortex and the known conduction velocities of horizontal fibers. The actual latencies increased with cortical distance, but much less than predicted by the horizontal propagation hypothesis. Probability calculations showed that an explanation of the context influence by lateral signal propagation is highly unlikely.

In contrast, mechanisms involving back projections from other extrastriate areas or subcortical structures (Craft et al., 2007; Jehee et al., 2007; Jeurissen et al., 2016) are plausible because they use white-matter fibers which are an order of magnitude faster than horizontal fibers. Context information for the 8 deg square that might take over 70 ms if conducted through horizontal fibers in V2 would take perhaps 10 ms if sent up to V4 and back.

Kogo et al. (2010), who base their model on perceptual observations of illusory figures akin to the Kanizsa triangle, state that "most of the many attempts to mimic the Kanizsa illusory phenomenon in neurocomputational models have been inspired by the borderline-completion scheme driven by the collinear alignment of the contours of the Pac Man shapes"—which is not true. In fact, all models since the mid 1980ies were inspired by the discovery of illusory contour responses in the visual cortex which included responses to stimuli that do *not* entail collinear alignment. When I began recording from area V2, I was surprised to find orientation selective neurons that responded to patterns consisting of lines orthogonal to their preferred orientation: lines that terminated along a virtual line through the receptive field at the preferred orientation (von der Heydt et al., 1984). Neurons that were sharply selective for a certain orientation responded vigorously to stimuli that had no line or edge of that orientation at all, and no energy for that orientation in the Fourier spectrum (von der Heydt and Peterhans, 1989). These stimuli also produce illusory contours in perception. A striking example of an illusory contour that is not a collinear completion of given features is the Ehrenstein illusion, in which a circular contour is produced by radial lines (Kogo et al. do not mention this illusion).

Also architects of artificial neural nets that do not claim physiological plausibility should take note that about 30% of the orientation selective cells in monkey V2 respond to a virtual line defined by line terminations as if it were a real line. V2 is a large area (in humans V2 is even larger than V1). Thus, 30% means a huge number of cells. There must be an advantage of having so many cells capable of signaling illusory contours. These cells seem to respond simply to the line of discontinuity, perhaps because it is indicative of an occluding contour. Their responses grow with the number of aligned terminations, but they do not require evidence for border ownership—the stimulus can be symmetric about the contour and does not need to have a closed contour or something that suggests a figure. V2 is an early stage in the process, and those responses appear with short latency.

Heitger et al. (1998) modeled the illusory contour neurons by combining two inputs, one that detects edges, and a second input that integrates termination features along the receptive field axis. They suggested that termination features are signaled by end-stopped cells (Heitger et al., 1992). Indeed, the neural illusory-contour responses had opened eyes for an important role of orthogonal features in the definition of contours. This model reproduced all the neural illusory contour responses and also produced the circular shapes of the Ehrenstein illusion. It achieved all this with a semi-local image operator.

As explained above, Craft et al. (2007) showed that integrating co-circular edge signals alone is not sufficient to reproduce the neural border ownership signals in configurations of partially occluding figures, and therefore included integration of T-junction signals, and von der Heydt and Zhang (2018) explicitly showed the influence of contextual T-junctions, L-junctions, and orthogonal edges in modulating the neural responses. Craft et al. adopted the two-input scheme of Heitger et al. (1992) and showed that it explains the data on neural responses to geometrical figures completely. I think there are good reasons to expect that an image-computable model that combines integration of co-circular edge signals as in Hu et al. (2019a) with integration of end-stopped signals as in Heitger et al. would improve the consistency of border ownership assignment, perhaps from the 69% score of Hu et al. to over 90%, as found in some neurons.

The notion that border ownership coding appears at low levels of the hierarchy and early in the process runs counter to current trends in machine vision. In convolutional nets one expects such context-sensitive coding only at higher levels, and late in the process. In fact, Hu et al. (2019b) found that the convolutional nets that represent figure-ground organization show it only at the higher levels.

Outlook

As said, area V4 is but one of many candidate regions in the search for grouping cells. In a way, the negative result in this visual area makes sense because representing objectness may require comprehensive action at multiple cortical levels. In fact, border ownership modulates responses in V1, V2, and V4, and shape selectivity of neurons in infero-temporal cortex also depends on border ownership. And for effective object-based attention, grouping cells should target neurons not only in V2, but at various levels of the visual object processing pathways in parallel, including V1, V2, V4, and IT. Indeed, recordings from different levels of the visual pathways have shown that attentional modulation tends to get stronger at higher levels, suggesting that the modulatory effects accumulate from stage to stage. Thus, grouping cells might not be found within the feature processing visual pathways, but rather in a structure "on the side" as sketched in Figure 8 (a similar architecture was proposed by Wolfe and Horowitz, 2004 for guidance in visual search, suggesting that "the 'guiding representation' … is not, itself, part of the pathway"). This idea also explains the finding that border ownership signals in V4 have similar or even shorter latencies than those of V2 (Bushnell et al., 2011; Franken and Reynolds, 2021).

Moreover, for dealing with objects, grouping cells should also receive and target neurons in other modalities like touch, proprioception, audition, taste, and smell. The model sketched in Figure 8 could be extended across modalities. The feed forward pathway activates grouping cells which provides handles for selective attention: The sound of a dropping coin directs visual attention to the site where the coin fell. Through back projections, grouping cells facilitate feature signals for the computation of object attributes: Say, an object has been identified visually. When the hand grasps the object, grouping cells selectively facilitate feature signals from skin and tendon receptors informing about haptic qualities and hand conformation, signals from which further processing may compute shape, weight, and other attributes of the object.

An important function of grouping cells and object pointers is in representing the layout of objects in a scene for reaching. When we reach for a pawn on a chess board, the hand easily grasps the pawn without knocking over other pieces on the board. This cannot be based on object recognition—all pawns look the same. Also selectively attending just to the target would not be successful. Grouping cells indicate object locations in retinal space, and object pointers track their locations in real space thus representing the layout of the objects.

Considering all these aspects it becomes clear that object representation needs a brain structure bigger than area V4. It must be large enough to be able to coordinate spatial information coming in through senses as diverse as vision, audition, and touch. Auditory space sense depends on head orientation, and so does vision, with the extra complication of eye movements, and tactile perception of 3D objects involves hand conformation. To combine these requires massive computations in real time. And we are looking for a structure that has connections to a range of cortical areas.

The pulvinar might be able to meet these requirements. The pulvinar is enlarged in primates which use hands for grasping and handling objects, compared to rats and cats which lack hands. It synchronizes activity between interconnected cortical areas according to attentional allocation (Saalmann et al., 2012). In humans, damage to the pulvinar often produces neglect (Ohye, 2002; Furman, 2014) which suggests a deficiency of grouping cells because grouping cells provide objects with "handles" for selective attention, and without these handles the system may not be able to disentangle objects in the visual representations even though the feature representations are intact.

The deficits expected from a loss of grouping cells are subtle; problems with visual attention to objects, visual guidance of grasping movements and saccades in cluttered scenes, e.g., situations where objects are partially occluded. Also deficits in object permanence and in maintaining object identity across object movements and saccades are to be expected.

Clearly, using object permanence as the criterion in the search for grouping cells is but one of many possible strategies. But it seems to me that permanence is the most decisive evidence for object-based perceptual organization. Grouping cells are a hypothesis of modeling, and a computational model is merely an existence proof. It shows that an algorithm exists that can perform a given task. Whether such cells really exist we do not know, they are imaginary. But persistence of border ownership signals is real.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Acknowledgments

I wish to thank Ernst Niebur for complementing my neurophysiology with computational neuroscience; Fangtu T. Qiu for creating a powerful and versatile system for visual stimulus generation, behavioral control, and recording; and Ofelia Garalde who as an animal lab technician contributed a lot to the experimental success of the 14 neurophysiological studies reviewed here.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Baylis, G. C., and Driver, J. (2001). Shape-coding in IT cells generalizes over contrast and mirror reversal, but not figure-ground reversal. *Nat. Neurosci.* 4, 937–942. doi: 10.1038/nn0901-937

Brincat, S. L., and Connor, C. E. (2006). Dynamic shape synthesis in posterior inferotemporal cortex. *Neuron* 49, 17–24. doi: 10.1016/j.neuron.2005.11.026

Bullier, J., Hupe, J. M., James, A. C., and Girard, P. (2001). the role of feedback connections in shaping the responses of visual cortical neurons. *Prog. Brain Res.* 134, 193–204. doi: 10.1016/S0079-6123(01)34014-1

Bushnell, B. N., Harding, P. J., Kosai, Y., and Pasupathy, A. (2011). Partial occlusion modulates contour-based shape encoding in primate area V4. *J. Neurosci.* 31, 4012–4024. doi: 10.1523/JNEUROSCI.4766-10.2011

Craft, E., Schütze, H., Niebur, E., and von der Heydt, R. (2007). A neural model of figure-ground organization. *J. Neurophysiol.* 97, 4310–4326. doi: 10.1152/jn.00203.2007

Franken, T., and Reynolds, J. (2021). Columnar processing of border ownership in primate visual cortex. *Elife* 10, e72573. doi: 10.7554/eLife.72573.sa2

Friedman, H. S., Zhou, H., and von der Heydt, R. (2003). The coding of uniform color figures in monkey visual cortex. *J. Physiol.* 548, 593–613. doi: 10.1113/jphysiol.2002.033555

Furman, M. (2014). "Chapter 19 - visual network," in *Neuronal Networks in Brain Function, CNS Disorders, and Therapeutics*, eds L. Carl Faingold, and H. Blumenfeld (San Diego, CA: Academic Press), 247–259.

Gillary, G., von der Heydt, R., and Niebur, E. (2017). Short-term depression and transient memory in sensory cortex. *J. Comput. Neurosci.* 43, 273–294. doi: 10.1007/s10827-017-0662-8

Grossberg, S., and Mingolla, E. (1985). Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading. *Psychol. Rev.* 92, 173–211. doi: 10.1037/0033-295X.92.2.173

Heitger, F., Rosenthaler, L., Von Der Heydt, R., Peterhans, E., and Kübler, O. (1992). Simulation of neural contour mechanisms: from simple to end-stopped cells. *Vis. Res.* 32, 963–981. doi: 10.1016/0042-6989(92)90039-L

Heitger, F., von der Heydt, R., Peterhans, E., Rosenthaler, L., and Kübler, O. (1998). Simulation of neural contour mechanisms: representing anomalous contours. *Image Vis. Comp. Comput. Psychophys. Stud. Early Vis.* 16, 407–421. doi: 10.1016/S0262-8856(97)00083-8

Hu, B., Khan, S., Niebur, E., and Tripp, B. (2019b). "Figure-ground representation in deep neural networks," in *2019 53rd Annual Conference on Information Sciences and Systems (CISS)* (Baltimore, MD), 1–6. doi: 10.1109/CISS.2019.8693039

Hu, B., von der Heydt, R., and Niebur, E. (2019a). Figure-ground organization in natural scenes: performance of a recurrent neural model compared with neurons of area V2. *ENeuro* 6, ENEURO.0479-18. doi: 10.1523/ENEURO.0479-18.2019

Intriligator, J., and Cavanagh, P. (2001). The spatial resolution of visual attention. *Cognit. Psychol* 43, 171–216. doi: 10.1006/cogp.2001.0755

Jehee, J. F., Lamme, V. A., and Roelfsema, P. R. (2007). Boundary assignment in a recurrent network architecture. *Vis. Res.* 47, 1153–1165. doi: 10.1016/j.visres.2006.12.018

Jeurissen, D., Self, M. W., and Roelfsema, P. R. (2016). Serial grouping of 2D-image regions with object-based attention in humans. Elife 5, e14320. doi: 10.7554/eLife.14320

Kogo, N., Strecha, C., Van Gool, L., and Wagemans, J. (2010). Surface construction by a 2-D differentiation-integration process: a neurocomputational model for perceived border ownership, depth, and lightness in Kanizsa Figures. *Psychol. Rev.* 117, 406–439. doi: 10.1037/a0019076

Lamme, V. A. F. (1995). The neurophysiology of figure-ground segregation in primary visual cortex. *J. Neurosci.* 15, 1605–1615. doi: 10.1523/JNEUROSCI.15-02-01605.1995

Martin, A. B., and von der Heydt, R. (2015). Spike synchrony reveals emergence of proto-objects in visual cortex. *J. Neurosci.* 35, 6860–6870. doi: 10.1523/JNEUROSCI.3590-14.2015

Martin, D., Fowlkes, C., Tal, D., and Malik, J. (2001). "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings of Eighth IEEE International Conference on Computer Vision* (Vancouver, BC: IEEE), 416–423. doi: 10.1109/ICCV.2001.937655

Mihalas, S., Dong, Y., von der Heydt, R., and Niebur, E. (2011). Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention to objects. *Proc. Nat. Acad. Sci. U. S. A.* 108, 7583–7588. doi: 10.1073/pnas.1014655108

Nakayama, K., Shimojo, S., and Silverman, G. H. (1989). Stereoscopic depth: its relation to image segmentation, grouping, and the recognition of occluded objects. *Perception* 18, 55–68. doi: 10.1068/p180055

O'Herron, P., and von der Heydt, R. (2009). Short-term memory for figure-ground organization in the visual cortex. *Neuron* 61, 801–809. doi: 10.1016/j.neuron.2009.01.014

O'Herron, P., and von der Heydt, R. (2011). Representation of object continuity in the visual cortex. *J. Vis.* 11, 12. doi: 10.1167/11.2.12

O'Herron, P., and von der Heydt, R. (2013). Remapping of border ownership in the visual cortex. *J. Neurosci.* 33, 1964–1974. doi: 10.1523/JNEUROSCI.2797-12.2013

Ohye, C. (2002). "Thalamus and thalamic damage," in *Encyclopedia of the Human Brain*, eds V. S. Ramachandran (New York, NY: Academic Press), 575–597.

Qiu, F. T., Sugihara, T., and von der Heydt, R. (2007). Figure-ground mechanisms provide structure for selective attention. *Nat. Neurosci.* 10, 1492–1499. doi: 10.1038/nn1989

Qiu, F. T., and von der Heydt, R. (2005). Figure and ground in the visual cortex: v2 combines stereoscopic cues with gestalt rules. *Neuron* 47, 155–166. doi: 10.1016/j.neuron.2005.05.028

Qiu, F. T., and von der Heydt, R. (2007). Neural representation of transparent overlay. *Nat. Neurosci.* 10, 283–284. doi: 10.1038/nn1853

Roelfsema, P. R., Lamme, V. A., and Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature.* 395, 376–381. doi: 10.1038/26475

Saalmann, Y. B., Pinsk, M. A., Wang, L., Li, X., and Kastner, S. (2012). The pulvinar regulates information transmission between cortical areas based on attention demands. *Science* 337, 753–756. doi: 10.1126/science.1223082

Smith, M. A., and Kohn, A. (2008). Spatial and temporal scales of neuronal correlation in primary visual cortex. *J. Neurosci.* 28, 12591–12603. doi: 10.1523/JNEUROSCI.2929-08.2008

Sugihara, T., Qiu, F. T., and von der Heydt, R. (2011). The speed of context integration in the visual cortex. *J. Neurophysiol.* 106: 374–385. doi: 10.1152/jn.00928.2010

Ungerleider, L. G., Galkin, T. W., Desimone, R., and Gattass, R. (2008). Cortical connections of area V4 in the Macaque. *Cereb. Cortex* 18, 477–499. doi: 10.1093/cercor/bhm061

Ungerleider, L. G., and Mishkin, M. (1982). "Two cortical visual systems," in *Analysis of Visual Behavior*, eds D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield (Cambridge: MIT Press), 549–586.

von der Heydt, R. (2015). Figure-ground organization and the emergence of proto-objects in the visual cortex. *Front. Psychol.* 6, 1695. doi: 10.3389/fpsyg.2015.01695

von der Heydt, R., Macuda, T. J., and Qiu, F. T. (2005). Border-ownership dependent tilt aftereffect. *J. Opt. Soc. Am. Opt. A.* 22, 2222–2229. doi: 10.1364/JOSAA.22.002222

von der Heydt, R., and Peterhans, E. (1989). Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. *J. Neurosci.* 9, 1731–1748. doi: 10.1523/JNEUROSCI.09-05-01731.1989

von der Heydt, R., Peterhans, E., and Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science* 224, 1260–1262. doi: 10.1126/science.6539501

von der Heydt, R., Peterhans, E., and Dürsteler, M. R. (1992). Periodic-pattern-selective cells in monkey visual cortex. *J. Neurosci.* 12, 1416–1434. doi: 10.1523/JNEUROSCI.12-04-01416.1992

von der Heydt, R., Qiu, F. T., and He, Z. J. (2003). Neural mechanisms in border ownership assignment: motion parallax and gestalt cues. *J. Vis.* 3/9, 666. doi: 10.1167/3.9.666

von der Heydt, R., and Zhang, N. R. (2018). Figure and ground: how the visual cortex integrates local cues for global organization. *J. Neurophysiol.* 120, 3085–3098. doi: 10.1152/jn.00125.2018

von Holst, E., and Mittelstaedt, H. (1950). Das Reafferenzprinzip. Wechselwirkungen Zwischen Zentralnervensystem Und Peripherie. *Naturwissenschaften* 37, 464–476. doi: 10.1007/BF00622503

Wagatsuma, N., von der Heydt, R., and Niebur, E. (2016). Spike synchrony generated by modulatory common input through NMDA-type synapses. *J. Neurophysiol.* 116, 1418–1433. doi: 10.1152/jn.01142.2015

Williford, J. R., and von der Heydt, R. (2016a). Figure-ground organization in visual cortex for natural scenes. *ENeuro* 3, ENEURO.0127–0116. doi: 10.1523/ENEURO.0127-16.2016

Williford, J. R., and von der Heydt, R. (2016b). *Data Associated with Publication 'Figure-Ground Organization in Visual Cortex for Natural Scenes,' Version 1*. Johns Hopkins University Data Archive. doi: 10.7281/T1C8276W

Wolfe, J. M., and Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nat. Rev. Neurosci.* 5, 495–501. doi: 10.1038/nrn1411

Zhang, N. R., and von der Heydt, R. (2010). Analysis of the context integration mechanisms underlying figure-ground organization in the visual cortex. *J. Neurosci.* 30, 6482–6496. doi: 10.1523/JNEUROSCI.5168-0;9.2010

Zhaoping, L. (2005). Border ownership from intracortical interactions in visual area V2. *Neuron* 47, 147–153. doi: 10.1016/j.neuron.2005.04.005

Zhou, H., Friedman, H. S., and von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *J. Neurosci.* 20, 6594–6611. doi: 10.1523/JNEUROSCI.20-17-06594.2000

Zhu, S. D., Zhang, L. A., and von der Heydt, R. (2020). Searching for object pointers in the visual cortex. *J. Neurophysiol.* 123, 1979–1994. doi: 10.1152/jn.00112.2020

frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 29 June 2023
DOI 10.3389/fcomp.2023.1178450

OPEN ACCESS

EDITED BY
Dirk Bernhardt-Walther,
University of Toronto, Canada

REVIEWED BY
Stavros Tsogkas,
Samsung AI Center Toronto, Canada
Katherine Rebecca Storrs,
Justus Liebig University Giessen, Germany

*CORRESPONDENCE
Paria Mehrani
paria61@yorku.ca

RECEIVED 02 March 2023
ACCEPTED 12 June 2023
PUBLISHED 29 June 2023

CITATION
Mehrani P and Tsotsos JK (2023) Self-attention in vision transformers performs perceptual grouping, not attention.
Front. Comput. Sci. 5:1178450.
doi: 10.3389/fcomp.2023.1178450

Self-attention in vision transformers performs perceptual grouping, not attention

Paria Mehrani* and John K. Tsotsos

Department of Electrical Engineering and Computer Science, York University, Toronto, ON, Canada

Recently, a considerable number of studies in computer vision involve deep neural architectures called vision transformers. Visual processing in these models incorporates computational models that are claimed to implement attention mechanisms. Despite an increasing body of work that attempts to understand the role of attention mechanisms in vision transformers, their effect is largely unknown. Here, we asked if the attention mechanisms in vision transformers exhibit similar effects as those known in human visual attention. To answer this question, we revisited the attention formulation in these models and found that despite the name, computationally, these models perform a special class of relaxation labeling with similarity grouping effects. Additionally, whereas modern experimental findings reveal that human visual attention involves both feed-forward and feedback mechanisms, the purely feed-forward architecture of vision transformers suggests that attention in these models cannot have the same effects as those known in humans. To quantify these observations, we evaluated grouping performance in a family of vision transformers. Our results suggest that self-attention modules group figures in the stimuli based on similarity of visual features such as color. Also, in a singleton detection experiment as an instance of salient object detection, we studied if these models exhibit similar effects as those of feed-forward visual salience mechanisms thought to be utilized in human visual attention. We found that generally, the transformer-based attention modules assign more salience either to distractors or the ground, the opposite of both human and computational salience. Together, our study suggests that the mechanisms in vision transformers perform perceptual organization based on feature similarity and not attention.

KEYWORDS

vision transformers, attention, similarity grouping, singleton detection, odd-one-out

1. Introduction

The Gestalt principles of grouping suggest rules that explain the tendency of perceiving a unified whole rather than a mosaic pattern of parts. Gestaltists consider organizational preferences, or priors, such as symmetry, similarity, proximity, continuity and closure as grouping principles that contribute to the perception of a whole. These principles which rely on input factors and the configuration of parts can be viewed as biases that result in the automatic emergence of figure and ground. To Gestalt psychologists, the perceptual organization of visual input to figure and ground was an early stage of interpretation prior to processes such as object recognition and attention. In fact, they posited that higher-level processes operate upon the automatically emerged figure. Some proponents of emergent intelligence go as far as to undermine the effect of attention on perceptual organization. For example, Rubin, known for his face-vase illusion, presented a paper in 1926 titled "On the Non-Existence of Attention" (Berlyne, 1974).

Despite the traditional Gestalt view, modern experimental evidence suggests that in addition to low-level factors, higher-level contributions can affect figure-ground organization. Specifically, experimental findings suggest that attention is indeed real and among the higher-level factors that influence figure-ground assignment (Qiu et al., 2007; Poort et al., 2012) (see Peterson, 2015 for review). Considering these discoveries and the enormous literature on attention (see Itti et al., 2005, for example), an interesting development in recent years has been the introduction of deep neural architectures dubbed transformers that claim to incorporate attention mechanisms in their hierarchy (Vaswani et al., 2017). Transformers, originally introduced in the language domain, were "based solely on attention mechanisms, dispensing with recurrence and convolutions entirely" (Vaswani et al., 2017).

Following the success of transformers in the language domain, Dosovitskiy et al. (2021) introduced the vision transformer (ViT), a transformer model based on self-attention mechanisms that received a sequence of image patches as input tokens. Dosovitskiy et al. (2021) reported comparable performance of ViT to convolutional neural networks (CNNs) in image classification and concluded, similar to (Vaswani et al., 2017), that convolution is not necessary for vision tasks. The reported success of vision transformers prompted a myriad of studies (Bhojanapalli et al., 2021; Caron et al., 2021; Dai et al., 2021; D'Ascoli et al., 2021; Liu et al., 2021, 2022; Mahmood et al., 2021; Srinivas et al., 2021; Touvron et al., 2021; Wu B. et al., 2021; Wu H. et al., 2021; Xiao et al., 2021; Yang et al., 2021; Yuan et al., 2021; Zhou et al., 2021; Bao et al., 2022; Guo et al., 2022; Han et al., 2022; Pan et al., 2022; Park and Kim, 2022; Zhou D. et al., 2022). In most of these studies, the superior performance of vision transformers, their robustness (Bhojanapalli et al., 2021; Mahmood et al., 2021; Naseer et al., 2021) and more human-like image classification behavior compared to CNNs (Tuli et al., 2021) were attributed to the attention mechanisms in these architectures. Several hybrid models assigned distinct roles of feature extraction and global context integration to convolution and attention mechanisms, respectively, and reported improved performance over models with only convolution or attention (Dai et al., 2021; D'Ascoli et al., 2021; Srinivas et al., 2021; Wu B. et al., 2021; Wu H. et al., 2021; Xiao et al., 2021; Guo et al., 2022). Hence, these studies suggested the need for both convolution and attention in computer vision applications.

A more recent study by Zhou Q. et al. (2022), however, reported that hybrid convolution and attention models do not "have an absolute advantage" compared to pure convolution or attention-based neural networks when their performance in explaining neural activities of the human visual cortex from two neural datasets was studied. Similarly, Liu et al. (2022) questioned the claims on the role of attention modules in the superiority of vision transformers by proposing steps to "modernize" the standard ResNet (He et al., 2016) into a new convolution-based model called ConvNeXt. They demonstrated that ConvNeXt with no attention mechanisms achieved competitive performance to state-of-the-art vision transformers on a variety of vision tasks. This controversy on the necessity of the proposed mechanisms compared to convolution adds to the mystery of the self-attention modules in vision transformers. Surprisingly, and to the best of our knowledge, no previous work directly investigated whether the self-attention modules, as claimed, implement attention mechanisms with effects similar to those reported in humans. Instead, the conclusions in previous studies were grounded on the performance of vision transformers vs. CNNs on certain visual tasks. As a result, a question remains outstanding: Have we finally attained a deep computational vision model that explicitly integrates visual attention into its hierarchy?

Answering this question is particularly important for advances in both human and computer vision fields. Specifically, in human vision sciences, the term attention has a long history (e.g., Berlyne, 1974; Tsotsos et al., 2005) and entails much confusion (e.g., Di Lollo, 2018; Hommel et al., 2019; Anderson, 2023). In a review of a book on attention (Sutherland, 1998) says,"After many thousands of experiments, we know only marginally more about attention than about the interior of a black hole". More recently, Anderson (2023) calls attention a conceptually fragmented term, a term that is assumed to have one meaning is found to have many, and suggests aid from mathematical language for theoretical clarity. The call for a more formal approach to vision research has appeared several times (e.g., Zucker, 1981; Tsotsos, 2011; Anderson, 2023) but no broadly accepted specification of attention is available. The majority of words in any dictionary have multiple meanings, and a particular class of words, homonyms, are spelled and pronounced the same yet differ in meaning which is only distinguished by the context in which they are used. "Attention" is one such word, here we seek to understand the scope of its use in order to provide the correct context.

To complicate matters further, many kinds of visual attention have been identified, the primary distinctions perhaps being that of overt and covert attention (with and without eye movements and viewpoint changes, respectively). Tsotsos (2011) shows over 20 kinds in his taxonomy, and other comprehensive reviews on the topic such as Desimone and Duncan (1995), Pashler (1998), Kastner and Ungerleider (2000), Itti et al. (2005), Styles (2006), Knudsen (2007), Nobre et al. (2014), Moore and Zirnsak (2017), and Martinez-Trujillo (2022) similarly cover many kinds, not all the same. As Styles (2006) asserts, attention is not a unitary concept. In addition, discussions of attention are always accompanied by consideration of how attention can change focus; this dynamic aspect does not appear in transformers at all.

The many descriptions of attention often conflate mechanism with effect while assuming that an exposition within some narrow domain easily generalizes to all of cognitive behavior. One might think that as long as the discussion remains within a particular community, all can be controlled with respect to use of terminology. This is not the case. Machine learning approaches have been already employed frequently in recent years in brain research by utilizing deep neural architectures as mathematical models of the brain (Cadieu et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Kubilius et al., 2016; Eickenberg et al., 2017; Zhuang et al., 2021). Therefore, it is only a matter of time before vision transformers with attention modules are used in human vision studies, if not already by the time of this publication. As a result, it is imperative to understand how attention modules in vision transformers relate to attention mechanisms in the

human visual system to avoid adding further confusion to attention research in human vision sciences.

Similarly, on the computer vision side, a more engineering kind of discipline, we need to specify the requirements of a solution against which we test the results of any algorithm realization. But the requirements of attention modules in vision transformers are not specified. They are only implied, through the use of the term 'attention' and can be traced back to the studies that explicitly motivated these modules, specifically, by the effect of attention mechanisms in the human visual system (i.e.,Vaswani et al., 2017→ Kim et al., 2017→Xu et al., 2015).

One might argue that from an engineering point of view, there is no need for these modules to remain faithful to their biological counterparts, hence, there is no need for direct comparison between the two systems. However, that train has already left the station. Computer vision has been using the term "attention" since the mid-1970's, connected to both inspiration from and comparisons to human visual attention, and continuously to this day (there are many reviews as evidence, e.g., Tsotsos and Rothenstein, 2011; Borji and Itti, 2012; Bylinskii et al., 2015). An expectation that a new mechanism can affect amnesia for a whole field is unwarranted. For example, Tan et al. (2021), Yue et al. (2021), Zhu et al. (2021), Paul and Chen (2022), and Panaetov et al. (2023) among others, have already mentioned effects of these modules as similar to those of attention in the human visual system.

Regardless of whether one considers attention from a human vision perspective or a machine vision point of view, it is unprincipled to leave the term ill-defined. Our goal in this paper is to contribute to an understanding of the function of the attention modules in vision transformers by revisiting two of their aspects. First, we hope to show that transformers formulate attention according to similarity of representations between tokens, and that this results in perceptual similarity grouping, not any of the many kinds of attention in the literature. Second, because of their feed-forward architecture, vision transformers cannot be not affected by factors such as goals, motivations, or biases (also see Herzog and Clarke, 2014). Such factors have played a role in attention models in computer vision for decades. Vision Transformers fall into the realm of the traditional Gestalt view of automatic emergence of complex features.

In a set of experiments, we examined attention modules in various vision transformer models. Specifically, to quantify Gestalt-like similarity grouping, we introduced a grouping dataset of images with multiple shapes that shared/differed in various visual feature dimensions and measured grouping of figures in these architectures. Our results on a family of vision transformers indicate that the attention modules, as expected from the formulation, group image regions based on similarity. Our second observations indicates that if vision transformers implement attention, it can only be in the form of bottom-up attention mechanisms. To test this observation, we measured the performance of vision transformers in the task of singleton detection. Specifically, a model that implements attention is expected to almost immediately detect the pop-out, an item in the input that is visually distinct from the rest of the items. Our findings suggest that vision transformers perform poorly in that regard and even in comparison to CNN-based saliency algorithms.

To summarize, our observations and experimental results suggest that "attention mechanisms" is a misnomer for computations implemented in so-called self-attention modules of vision transformers. Specifically, these modules perform similarity grouping and not attention. In fact, the self-attention modules implement a special class of in-layer lateral interactions that were missing in CNNs (and perhaps this is the reason for their generally improved performance). Lateral interactions are known as mechanisms that counteract noise and ambiguity in the input signal (Zucker, 1978). In light of this observation, the reported properties of vision transformers such as smoothing of feature maps (Park and Kim, 2022) and robustness (Mahmood et al., 2021; Naseer et al., 2021) can be explained. These observations lead to the conclusion that the quest for a deep computational vision model that implements attention mechanisms has not come to an end yet.

In what follows, we will employ the terms attention and self-attention interchangeably as our focus is limited to vision transformers with transformer encoder blocks. Also, each computational component in a transformer block will be referred to as a module, for example, the attention module or the multi-layer perceptron (MLP) module. Both block and layer, then, will refer to a transformer encoder block that consists of a number of modules.

2. Materials and methods

In this section, we first provide a brief overview of vision transformers followed by revisiting attention formulation and the role of architecture in visual processing in these models. Then, we explain the details of the two experiments we performed in this study.

2.1. Vision transformers

Figure 1 provides an overview of Vision Transformer (ViT) and the various modules in its transformer encoder blocks. Most vision transformer models extend and modify or simply augment a ViT architecture into a larger system. Regardless, the overall architecture and computations in the later variants resemble those of ViT and each model consists of a number of stacked transformer encoder blocks. Each block performs visual processing of its input through self-attention, MLP and layer normalization modules. Input to these networks includes a sequence of processed image tokens (localized image patches) concatenated with a learnable class token.

Vision transformer variants can be grouped into three main categories:

1. Models that utilized stacks of transformer encoder blocks as introduced in ViT but modified the training regime and reported a boost in performance, such as DeiT (Touvron et al., 2021) and BEiT (Bao et al., 2022).
2. Models that modified ViT for better adaptation to the visual domain. For example, Liu et al. (2021) introduced an architecture called Swin and suggested incorporating various scales and shifted local windows between blocks. A few other

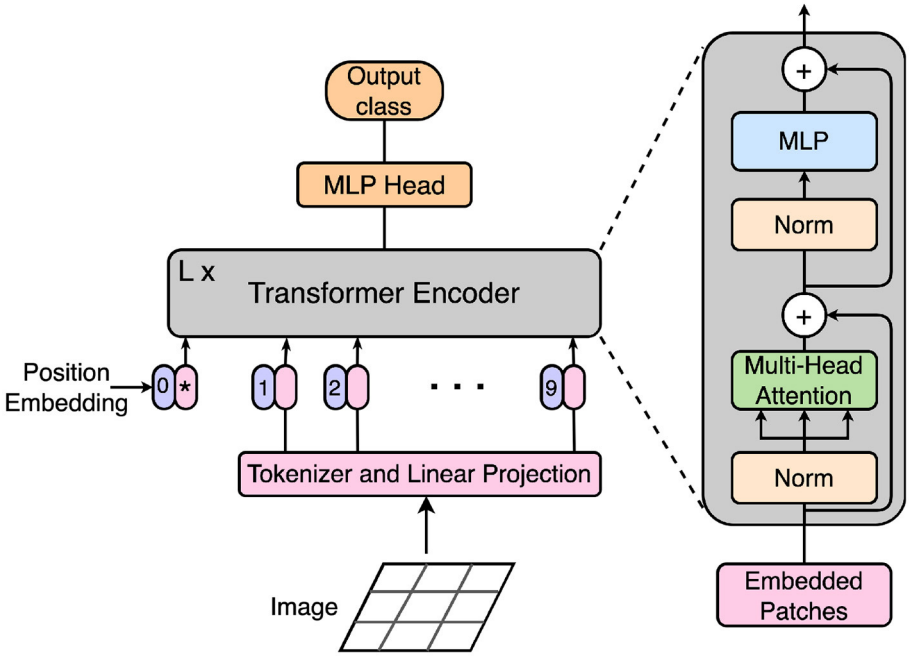


FIGURE 1
The ViT model architecture (Dosovitskiy et al., 2021). First, each input image is split into local patches called tokens. After linear embedding of the tokens, a numerical position embedding is added to each token. After concatenating a learnable class embedding shown with an asterisk to the input sequence, the combined embeddings are fed to L blocks of transformer encoders. The output of the final encoder block is fed to a classification head in ViT. The zoomed-in diagram on the right demonstrates the various modules within a transformer encoder block. These modules consist of norm, multi-head self-attention and MLP.

work suggested changes to the scope of attention, for example, local vs. global (Chen et al., 2021; Yang et al., 2021).

3. Hybrid models that introduced convolution either as a preprocessing stage (Xiao et al., 2021) or as a computational step within transformer blocks (Wu H. et al., 2021).

The family of vision transformers that we studied in our experiments includes ViT, DEiT, BEiT, Swin, and CvT. These models span all three categories of vision transformers as classified above. For each model, we studied a number of pre-trained architectures available on HuggingFace (Wolf et al., 2020). Details of these architectures are outlined in Table 1.

2.1.1. Attention formulation

In transformers, the attention mechanism for a query and key-value pair is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \quad (1)$$

where Q, K, and V represent matrices of queries, keys and values with tokens as rows of these matrices, and d_k is the dimension of individual key/query vectors. Multiplying each query token, a row of Q, in the matrix multiplication QK^T is in fact a dot-product of each query with all keys in K. The output of this dot-product can be interpreted as how similar the query token is to each of the key tokens in the input; a compatibility measure. This dot product is then scaled by $\sqrt{d_k}$ and the softmax yields the weights for value tokens. Vaswani et al. (2017) explained the output of attention modules as "a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key". The same formulation was employed in ViT while the compatibility function formulation is slightly modified in some vision transformer variants. Nonetheless, the core of the compatibility function in all of these models is a dot-product measuring representation similarity. Vaswani et al. (2017) reported improved performance when instead of a single attention function, they mapped the query, key and value tokens to h disjoint representational learned spaces and computed attention in each space called a head. Concatenation of the attention computed in individual heads yields the output of the attention module that they called Multi-Head Attention module.

In transformer encoders, the building block of vision transformers, the query, key and value have the same source and come from the output of the previous block. Hence, the attention modules in these blocks are called self-attention. In this case, the attention formulation can be explained as a process that results in consistent token representations across all spatial positions in the stimulus. Specifically, token representation and attention can be described as follows: each token representation signifies presence/absence of certain visual features, providing a visual interpretation or label at that spatial position. The attention mechanism incorporates the context from the input into its process and describes the inter-token relations determined by the compatibility function. As a result, Equation (1) shifts the interpretation of a given token toward that of more compatible

TABLE 1 The family of vision transformers studied in this work.

Model	Architecture name	# layers	# params	Training dataset	Fine-tuned
ViT	ViT-base-patch16-224	12	86 M	ImageNet-21k	–
DeiT	DeiT-tiny-distilled-patch16-224	12	5 M	ImageNet-1k	ImageNet-1k
	DeiT-small-distilled-patch16-224	12	22 M	ImageNet-1k	ImageNet-1k
	DeiT-base-distilled-patch16-224	12	86 M	ImageNet-1k	ImageNet-1k
BEiT	BEiT-base-patch16-224	12	86 M	ImageNet-21k	ImageNet-1k
	BEiT-base-patch16-224-pt22k	12	86 M	ImageNet-21k	–
	BEiT-base-patch16-224-pt22k-ft22k	12	86 M	ImageNet-21k	ImageNet-21k
CvT	CvT-13	13	19.98 M	ImageNet-1k	–
	CvT-21	21	31.54 M	ImageNet-1k	–
Swin	Swin-tiny-patch4-window7-224	12	29 M	ImageNet-1k	–
	Swin-small-patch4-window7-224	12	50 M	ImageNet-1k	–

For each model, a number of architecture variations were studied. For all models, pre-trained architectures available on HuggingFace (Wolf et al., 2020) were utilized. Input resolution to all pre-trained models was 224 × 224. The datasets used for training and fine-tuning are specified. Whereas, DeiT and BEiT models use the same general architecture as ViT, Swin introduces multiple scales and shifted windows to overcome the shortcomings of fixed size and position in tokens for visual tasks. The CvT architectures are hybrid models combining convolution and self-attention mechanisms in each transformer encoder block.

tokens in the input. The final outcome of this process will be groups of tokens with similar representations. Zucker (1978) referred to this process as "Gestalt-like similarity grouping process".

In Zucker (1978), the Gestalt-like similarity grouping process is introduced as a type of relaxation labeling (RL) process. Relaxation labeling is a computational framework for updating the possibility of a set of labels (or interpretations) for an object based on the current interpretations among neighboring objects. Updates in RL are performed according to a *compatibility function* between labels. In the context of vision transformers, at a given layer, each token is an object for which a feature representation (label) is provided from the output of the previous layer. A token representation is then updated (the residual operation after the attention module) according to a dot-product compatibility function defined between representations of neighboring tokens. In ViT, the entire stimulus forms the neighborhood for each token.

Zucker (1978) defined two types of RL processes in low-level vision: vertical and horizontal. In horizontal processes, the compatibility function defines interaction at a single level of abstraction but over multiple spatial positions. In contrast, vertical processes involve interaction in a single spatial position but across various levels of abstraction. Although Zucker counts both types of vertical and horizontal processes contributing to Gestalt-like similarity grouping, self-attention formulation only fits the definition of horizontal relaxation labeling process and thus, implements a special class of RL. As a final note, while traditional RL relies on several iterations to achieve consistent labeling across all positions, horizontal processes in vision transformers are limited to a single iteration and therefore, a single iteration of Gestalt-like similarity grouping is performed in each transformer encoder block.

2.1.2. Transformer encoders are feed-forward models

Even though the formulation of self-attention in vision transformers suggests Gestalt-like similarity grouping, this alone does not rule out the possibility of performing attention in these modules. We consider this possibility in this section.

It is now established that humans employ a set of mechanisms, called visual attention, that limit visual processing to sub-regions of the input to manage the computational intractability of the vision problem (Tsotsos, 1990, 2017). Despite the traditional Gestalt view, modern attention research findings suggest a set of bottom-up and top-down mechanisms determine the target of attention. For example, visual salience ["the distinct subjective perceptual quality which makes some items in the world stand out from their neighbors and immediately grab our attention" (Itti, 2007)] is believed to be a bottom-up and stimulus-driven mechanism employed by the visual system to select a sub-region of the input for further complex processing. Purely feed-forward (also called bottom-up) processes, however, were shown to be facing an intractable problem with exponential computational complexity (Tsotsos, 2011). Additionally, experimental evidence suggests that visual salience (Desimone and Duncan, 1995) as well as other low-level visual factors could be affected by feedback (also known as top-down) and task-specific signals (Folk et al., 1992; Bacon and Egeth, 1994; Kim and Cave, 1999; Yantis and Egeth, 1999; Lamy et al., 2003; Connor et al., 2004; Baluch and Itti, 2011; Peterson, 2015). In other words, theoretical and experimental findings portray an important role for top-down and guided visual processing. Finally, Herzog and Clarke (2014) showed how a visual processing strategy for human vision cannot be both hierarchical and strictly feed-forward through an argument that highlights the role of visual context. A literature going back to the 1800's extensively documents human attentional abilities (Itti et al., 2005; Carrasco, 2011; Nobre et al., 2014; Tsotsos, 2022; Krauzlis et al., 2023).

Modern understanding of visual attention in humans provides a guideline to evaluate current computational models for visual attention. Vision transformers are among more recent developments that are claimed to implement attention mechanisms. However, it is evident that these models with their purely feed-forward architectures implement bottom-up mechanisms. Therefore, if it can be established that these models

implement attention mechanisms, they can only capture the bottom-up signals that contribute to visual attention and not all aspects of visual attention known in humans. These observations call for a careful investigation of the effect of attention on visual processing in these models.

2.2. Experiments

In our experiments, we will consider the output of the attention module in each model block (the green rectangle in Figure 1) before the residual connection. In both experiments, we removed the class token from our analysis. Suppose that an attention module receives an input of size $H \times W \times C$, where H, W, and C represent height, width and feature channels. Then, the output, regardless of whether the attention module is multi-head or not, will also be of size $H \times W \times C$. In what follows, the term attention map is used for each $H \times W$ component of the attention module output along each single feature dimension $c \in \{1, 2, \ldots, C\}$. In other words, the values comprising each attention map are obtained from the attention scores (Equation 1), along a single feature dimension. Also, feature channel and hidden channel will be employed interchangeably.

It is important to emphasize that the attention maps we consider for our experiments and evaluations differ from those often visualized in the vision transformer literature. Specifically, in our evaluations, we consider what the model deems as salient, the regions that affect further processing in later model blocks. In contrast, what is commonly called an attention map in previous work (Dosovitskiy et al., 2021) is computed for a token, usually the output token in vision transformers and by recursively backtracking the compatibility of the token with other tokens to the input layer (Abnar and Zuidema, 2020). Therefore, a different map can be plotted for the various class tokens in the model and these maps are conditioned on the given token. One can interpret these maps as regions of input that are most relevant to yielding the given class token. Also, note that the compatibility (result of the softmax function in Equation 1) employed for this visualization, is only part of what (Vaswani et al., 2017) called the attention score defined Equation 1. Maps obtained with this approach do not serve our goal: we seek to determine regions of the input that were considered as salient, as Xu et al. (2015) put it, and were the focus of attention during the bottom-up flow of the signal in inference mode. These regions with high attention scores from Equation (1) are those that affect the visual signal through the residual connection (the + sign after the green rectangle in Figure 1). Hence, we evaluated *the output of the attention module* in both experiments.

2.2.1. Experiment 1: similarity grouping

To quantify Gestalt-like similarity grouping in vision transformers, we created a dataset for similarity grouping with examples shown in Figure 2 and measured similarity grouping performance in vision transformers mentioned in Section 2.1. As explained earlier, the attention from Equation (1) signals grouping among tokens. Therefore, we measured similarity grouping by recording and analyzing the output of attention modules in these models.

2.2.1.1. Dataset

Each stimulus in the dataset consists of four rows of figures with features that differ along a single visual feature dimension including hue, orientation, lightness, shape, orientation and size. Each stimulus is 224 × 224 pixels and contains two perceptual groups of figures that alternate between the four rows. The values of the visual feature that formed the two groups in each stimulus were randomly picked.

In some vision transformers, such as ViT, the token size and position are fixed from input and across the hierarchy. This property has been considered a shortcoming in these models when employed in visual tasks and various work attempted to address this issue (Liu et al., 2021). Since we included vision transformers that employ ViT as their base architecture in our study, and in order to control for the token size and position in our analysis, we created the dataset such that each figure in the stimulus would fit within a single token of ViT. In this case, each figure fits a 16 × 16 pixels square positioned within ViT tokens. To measure the effect of fixed tokens on grouping, we created two other sets of stimuli. In the first set, we considered the center of every other token from ViT as a fixed position for figures and generated stimuli with figures that would fit 32 × 32 pixels squares. In this case, each figure will be relatively centered at a ViT token, but will span more than a single token. In the second set, we generated stimuli with figures that were token-agnostic. We designed these stimuli such that the set of figures was positioned at the center of the image instead of matching token positions, with each figure size fitting a 37 × 37 pixels square.

Each version of our grouping dataset consists of 600 images with 100 stimuli per visual feature dimension, summing to a total of 1,800 stimuli for all three versions.

2.2.1.2. Evaluation and metrics

For a self-attention module that yields a $H \times W \times C$ map, where H and W represent height and width and C the number of feature channels, we first normalized the attention maps across individual feature channels so that attention scores are in the $[0, 1]$ range. Then, we measured grouping along each feature channel based on two metrics:

- *Grouping index:* Suppose A_{g1} and A_{g2} represent the average attention score of pixels belonging to figures in group 1 and group 2, respectively. We defined the grouping index as:

$$GI = \frac{\|A_{g1} - A_{g2}\|}{A_{g1} + A_{g2}}. \tag{2}$$

  The grouping index GI varies in $[0, 1]$, with larger values indicating better grouping of one group of figures in the stimulus along the feature channel.
- *Figure-background ratio:* The overall performance of vision transformers will be impacted if background tokens are grouped with figure tokens (mixing of figure and ground). Therefore, we measured the figure-background attention ratio as:

$$AR = \max\left(\frac{A_{g1}}{A_{bkg}}, \frac{A_{g2}}{A_{bkg}}\right), \tag{3}$$

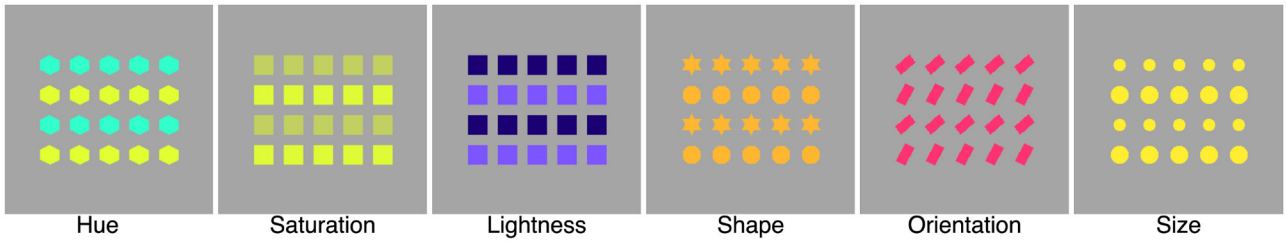


FIGURE 2
Similarity grouping stimuli examples. The stimuli in this dataset consists of two groups defined according to difference in one of hue, saturation, lightness, shape, orientation, size features. Each stimuli has four rows with alternating figures from the two groups. The values for the visual features that define the groups are chosen randomly. The shape set in this dataset consists of rectangle, triangle, ellipse, star, rhombus, right triangles, trapezoid, hexagon and square. Examples in this figure were picked from the version in which each figure fits within a 37 × 37 square.

where A_{g1}, A_{g2} represent the average attention for group 1 and group 2 figures, respectively, and A_{bkg} is the average score of background. The attention ratio AR is positive and values larger than 1 indicate the attention score of at least one group of figures is larger than that of the background (the larger the ratio, the less the mixing of figure and ground). Note that the attention ratio AR signifies the relative attention score assigned to figure and ground. Therefore, values close to 1 suggest similar attention scores assigned to figure and ground, quite contrary to the expected effect from attention mechanisms.

For each stimulus, we excluded all feature dimensions along which both $A_{g1} = 0$ and $A_{g2} = 0$ from our analysis. This happens when, for example, the feature channels represent green hues, and the figures in the stimulus are figures of red and blue. Moreover, when analyzing AR, we excluded all channels with $A_{bkg} = 0$ as our goal was to investigate grouping of figure and ground when some attention was assigned to the background.

2.2.2. Experiment 2: singleton detection

Evidence for similarity grouping does not disprove implementation of attention in vision transformers. Since these models are feed-forward architectures, investigating the effect of attention modules in their visual processing must be restricted to bottom-up mechanisms of attention. Therefore, we limited our study to evaluating the performance of these models in the task of singleton detection as an instance of saliency detection (see Bruce et al., 2015; Kotseruba et al., 2019 for a summary of saliency research). Specifically, strong performance on saliency detection would suggest that these models implement the bottom-up mechanisms deployed in visual attention.

In this experiment, we recorded the attention map of all blocks in vision transformers mentioned in Section 2.1. Following Zhang and Sclaroff (2013), we computed an average attention map for each transformer block by averaging over all the attention channels and considered the resulting map as a saliency map. Then, we tested if the saliency map highlights the visually salient singleton. Additionally, we combined the feature maps obtained after the residual operation of attention modules and evaluated saliency detection performance for the average feature map. It is worth noting that self-attention modules, and not the features maps, are expected to highlight salient regions as the next targets for further visual processing. Nonetheless, for a better understanding of the various representations and mechanisms in vision transformers, we included feature-based saliency maps in our study.

2.2.2.1. Dataset

For the singleton detection experiment, we utilized the psychophysical patterns (P^3) and odd-one-out (O^3) dataset introduced by Kotseruba et al. (2019). Examples of each set are shown in Figure 3. The P^3 dataset consists of 2,514 images of size 1,024×1,024. Each image consists of figures on a regular 7 × 7 grid with one item as the target that is visually different in one of color, orientation or size from other items in the stimulus. The location of the target is chosen randomly. The O^3 dataset includes 2,001 images with the largest dimension set to 1,024. In contrast to the grouping and P^3 datasets whose stimuli were synthetic images, the O^3 dataset consists of natural images. Each image captures a group of objects that belong to the same category with one that stands out (target) from the rest (distractors) in one or more visual feature dimensions (color, texture, shape, size, orientation, focus and location). The O^3 with natural images provides the opportunity to investigate the performance of the vision transformer models in this study on the same type of stimuli those were trained. Both P^3 and O^3 datasets are publicly available and further details of both datasets can be found in Kotseruba et al. (2019).

2.2.2.2. Metrics

We followed Kotseruba et al. (2019) to measure singleton detection performance in vision transformers. We employed their publicly available code for the computation of metrics they used to study traditional and deep saliency models. The number of fixation and saliency ratio were measured for P^3 and O^3 images, respectively, as explained below.

- *Number of fixations:* Kotseruba et al. (2019) used the number of fixations required to detect pop-out as a proxy for salience. Specifically, they iterated through the maxima of the saliency map until the target was detected or a maximum number of iterations was reached. At each iteration that resembles a fixation of the visual model on a region of input, they suppressed the fixated region with a circular mask before moving the fixation to the next maxima. Lower number of

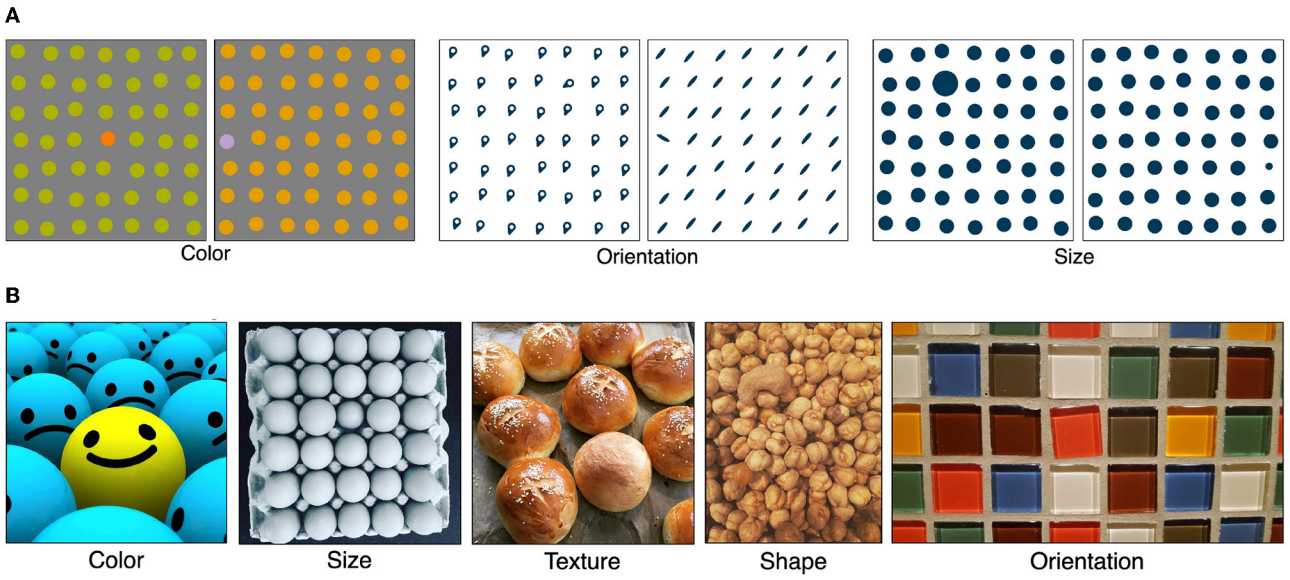


FIGURE 3
Samples of stimuli from P^3 and O^3 datasets introduced by Kotseruba et al. (2019) are illustrated. These datasets consist of stimuli with singletons in various feature dimensions. **(A)** Examples from the psychophysical patterns (P^3) dataset. The singletons in this dataset are defined according to color, orientation and size. **(B)** Examples from the odd-one-out (O^3) dataset with singletons in color, size, texture, shape and orientation feature dimensions.

fixations indicates higher relative salience of the target to that of distractors.

- *Saliency ratio:* Kotseruba et al. (2019) employed the ratio of the maximum saliency of the target vs. the maximum saliency of the distractors. They also measured the ratio of the maximum saliency of the background to the maximum saliency of the target. These two ratios that are referred to as MSR_{targ} and MSR_{bg} determine if the target is more salient than the distractors or the background, respectively. Ideally, MSR_{targ} is >1 and MSR_{bg} is <1.

3. Results

3.1. Experiment 1: similarity grouping

Each vision transformer in our study consists of a stack of transformer encoder blocks. In this experiment, our goal was to investigate similarity grouping in attention modules in transformer encoder blocks. We were also interested in changes in similarity grouping over the hierarchy of transformer encoders. Therefore, for each vision transformer, we took the following steps: We first isolated transformer encoders in the model and computed the grouping index (*GI*) and attention ratio (*AR*) per channel as explained in Section 2.2.1.2. Then, we considered the mean *GI* and *AR* per block as the representative index and ratio of the layer.

Figure 4A shows the mean *GI* for the architecture called "ViT-base-patch16-224" in Table 1 over all layers of the hierarchy. The *GI* is plotted separately according to the visual feature that differed between the groups of figures. This plot demonstrates that *GI* for all blocks of this model across all tested feature dimensions is distinctly larger than 0, suggesting similarity grouping of figures in all attention modules of this architecture. Interestingly, despite some variations in the first block, all layers have relatively similar *GI*. Moreover, the grouping indices for all feature dimensions are close, except for hue with *GI* larger than 0.6 in the first block, indicating stronger grouping among tokens based on this visual feature.

Figure 4B depicts the mean *AR* for the same architecture, ViT-base-patch16-224, for all the encoder blocks. Note that all curves in this plot are above the $AR = 1$ line denoted as a dashed gray line, indicating that all attention modules assign larger attention scores to at least one group of figures in the input vs. the background tokens. However, notable is the steep decline in the mean *AR* across the hierarchy. This observation confirms the previous reports of smoother attention maps in higher stages of the hierarchy (Park and Kim, 2022) with similar attention assigned to figure and background tokens.

Figure 5 shows the mean *GI* for all the architectures from Table 1 separately based on the visual feature that defined the groups in the input. All models, across all their layers, with some exceptions, demonstrate mean *GI* that are distinctly larger than 0. The exceptions include the first layer of all BEiT architectures and Swin-small-patch4-window7-224, and the last block of CvT-13 and CvT-21. Interestingly, BEiT and Swin architectures jump in their mean *GI* in their second block. Even though DeiT and BEiT architectures utilized the same architecture as ViT but trained the model with more modern training regimes, both models demonstrate modest improvement over ViT-base-patch16-224.

Plots in Figure 6 depict the mean *AR* over all the architectures. Interestingly, ViT-base-patch16-224 is the only architecture whose mean *AR* for the first block is the largest in its hierarchy and unanimously for all visual features. Among the three DeiT architectures (tiny, small, and base), DeiT-tiny-distilled-patch16-224, demonstrates larger mean AR

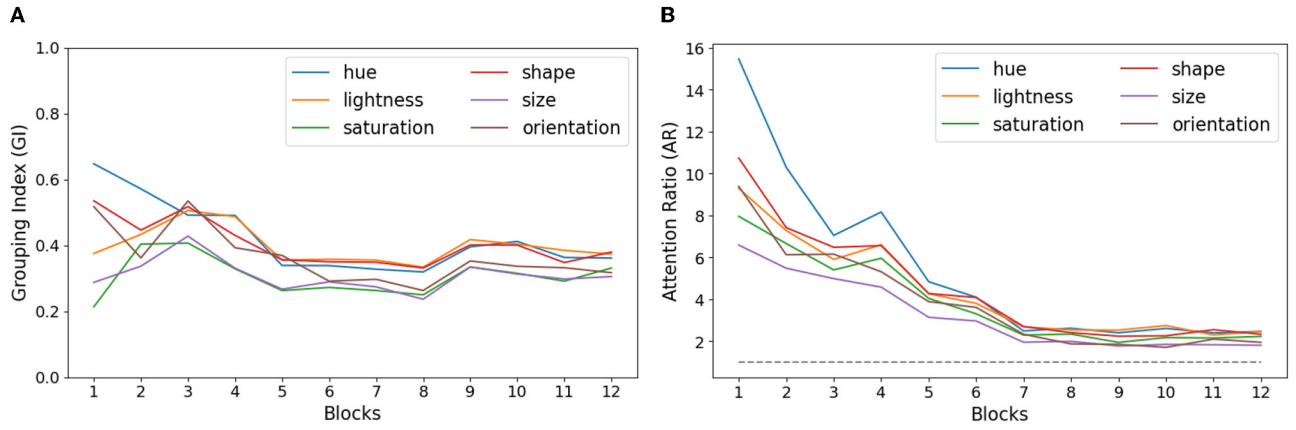


FIGURE 4
Mean grouping index and attention ratio for the ViT-base-patch16-224 architecture over all stimuli but separated according to the visual features that defined the groups of figures in the input. **(A)** The mean grouping index is larger than 0.2 for all layers of the model across all visual features, suggesting perceptual grouping based on similarity in this architecture. **(B)** The attention ratio of larger than 1 for all transformer encoder blocks of ViT-base-patch16-224 indicates larger scores are assigned to figure tokens. However, the steep decline in the *AR* ratio in the hierarchy demonstrates mixing of figure and background tokens due to similar attention scores. **(A)** Mean grouping index (GI). **(B)** Mean attention ratio (AR).

ratios. Comp ared to ViT, DeiT-tiny-distilled-patch16-224 has far fewer parameters and the comparable mean AR for this architecture with Vit confirms the suggestion of Touvron et al. (2021) that an efficient training regime in a smaller model could result in performance gain against a larger model. Results from Figure 6 are also interesting in that all of Swin and CvT architectures that are claimed to adapt transformer models to the vision domain, have relatively small mean *AR* over their hierarchy. These results show that these models mix figure and background tokens in their attention score assignments, an observation that deserves further investigation in a future work.

Finally, Figure 7 summarizes the mean grouping index *GI* for the DeiT-base-distilled-patch16-224 architecture over the three versions of the grouping dataset as explained in Section 2.2.1.1. These results demonstrate similar grouping index over all three versions, suggesting little impact of token position and size relative to figures in the input.

3.2. Experiment 2: singleton detection

Generally, in saliency experiments, the output of the model is considered for performance evaluation. In this study, however, not only we were interested in the overall performance of vision transformers (the output of the last block), but also in the transformation of the saliency signal in the hierarchy of these models. Examining the saliency signal over the hierarchy of transformer blocks would provide valuable insights into the role of attention modules in saliency detection. Therefore, we measured saliency detection in all transformer blocks.

3.2.1. The P^3 dataset results

Following Kotseruba et al. (2019), to evaluate the performance of vision transformer models on the P^3 dataset, we measured the target detection rate at 15, 25, 50, and 100 fixations. Chance level performance for ViT-base-patch16-224, as an example, would be 6, 10, 20, and 40% for 15, 25, 50, and 100 fixations, respectively (masking after each fixation explained in Section 2.2.2 masks an entire token). Although these levels for the various models would differ due to differences in token sizes and incorporating multiple scales, these chance level performances from ViT-base-patch16-224 give a baseline for comparison.

Figure 8 demonstrates the performance of saliency maps obtained from attention and feature maps of all ViT-base-patch16-224 blocks. These plots clearly demonstrate that the feature-based saliency maps in each block outperform those computed from the attention maps. This is somewhat surprising since as explained in Section 2.2.2, if vision transformers implement attention mechanisms, attention modules in these models are expected to highlight salient regions in the input for further visual processing. Nonetheless, plots in Figure 8 tell a different story, namely that feature maps are preferred options for applications that require singleton detection. Comparing target detection rates across color, orientation and size for both attention and feature maps demonstrate higher rates in detecting color targets compared to size and orientation. For all three of color, orientation and size, the target detection rates peak at earlier blocks for attention-based saliency maps and decline in later blocks, with lower than chance performance for most blocks. This pattern is somewhat repeated in feature-based saliency maps with more flat curves in the hierarchy, especially for a larger number of fixations.

Similar detection rate patterns were observed in other vision transformer models. However, due to limited space, we refrain from reporting the same plots as in Figure 8 for all the vision transformer models that we studied. These plots can be found in the Supplementary material. Here, for each model, we report the mean target detection rate over all blocks and the detection rate for the last block of each model for both attention and feature-based saliency maps. These results are summarized in Figures 9, 10 for the last and mean layer target detection rates, respectively. Consistent with the observations from ViT-base-patch16-224 in

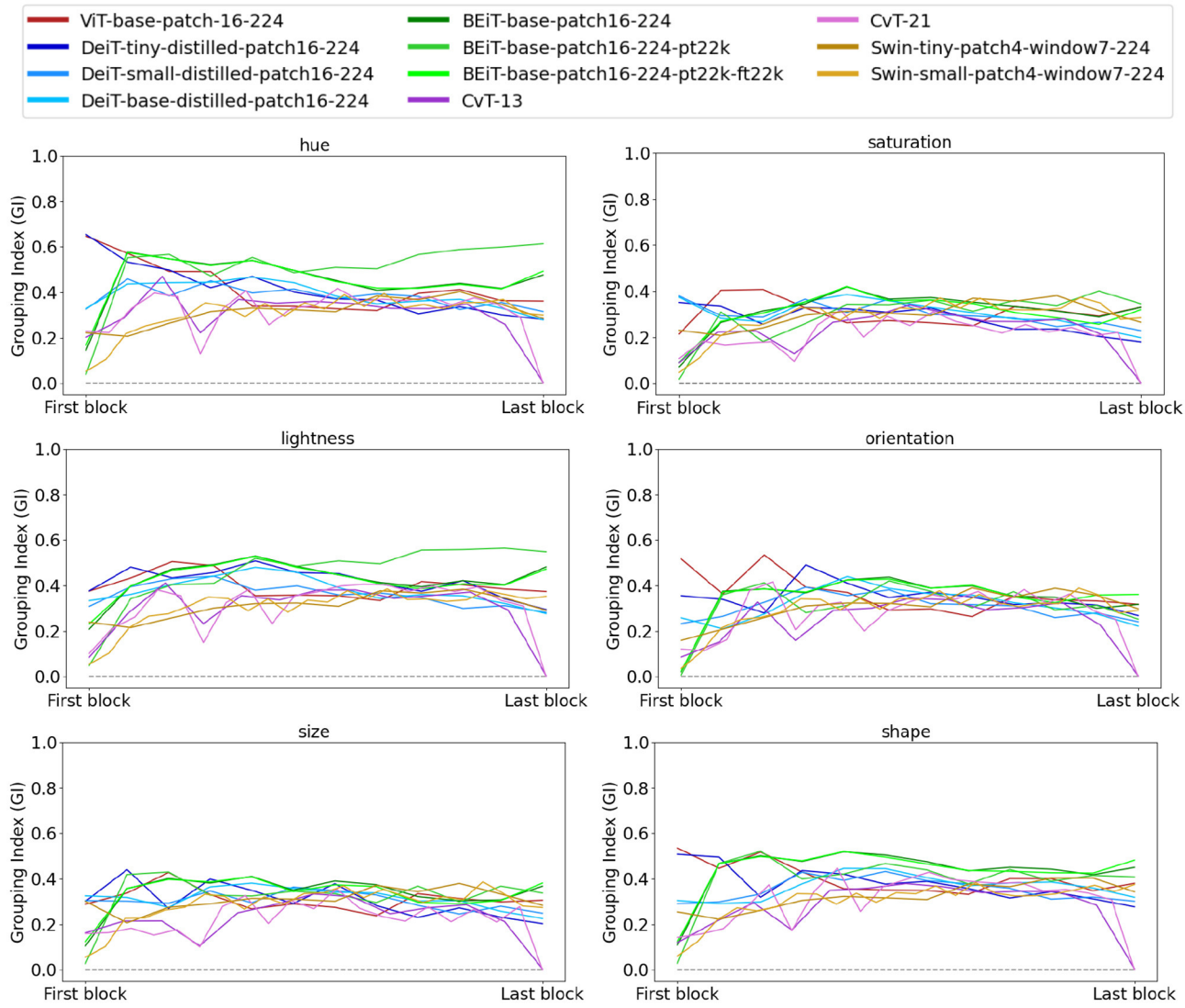


FIGURE 5
Mean grouping index for all the architectures from Table 1 plotted separately for each visual feature that defined the perceptual grouping of figures in the input. For comparison purposes, we plotted the grouping index of all models within the same range on the *x*-axis. Accordingly, we labeled the *first* and *last* block on the horizontal axis. These plots demonstrate that all models perform similarity grouping of figures based on low-level visual features such as hue and orientation. Except for the final block of CvT models, all layers of all architectures have mean *GI* higher than 0. The legend at the top applies to all plots.

Figure 8, the feature-based saliency maps outperform attention-based ones in Figure 9 and in general have higher detection rates than the chance levels stated earlier. The attention-based saliency maps, across most of the models, fail to perform better than chance. Generally, all models have higher detection rates for color targets, repeating similar results reported by Kotseruba et al. (2019). Interestingly, Swin architectures that incorporate multiple token scales, perform poorly in detecting size targets with both feature and attention-based saliency maps.

Results for mean target detection rates over all blocks in Figure 10 are comparable to those of last layer detection rates, except for a shift to higher rates. Specifically, all models are more competent at detecting color targets and that the feature-based saliency maps look more appropriate for singleton detection. In Swin architectures, the mean detection rate of feature-based saliency maps are relatively higher for size targets than that of other models. This observation, together with the last layer detection rate of Swin models for size targets suggest that incorporating multiple scales in vision transformers improves representing figures of various sizes but the effect fades higher in the hierarchy.

In summary, the attention maps in vision transformers were expected to reveal high salience for the target vs. distractors. Nonetheless, comparing the detection rate of attention-based saliency maps in vision transformers at 100 fixations with those of traditional and deep saliency models reported by Kotseruba et al. (2019) suggest that not only do the attention modules in vision transformers fail to highlight the target, but also come short of convolution-based deep saliency models with no attention modules. Although the feature-based saliency maps in vision transformers showed promising results in target detection rates

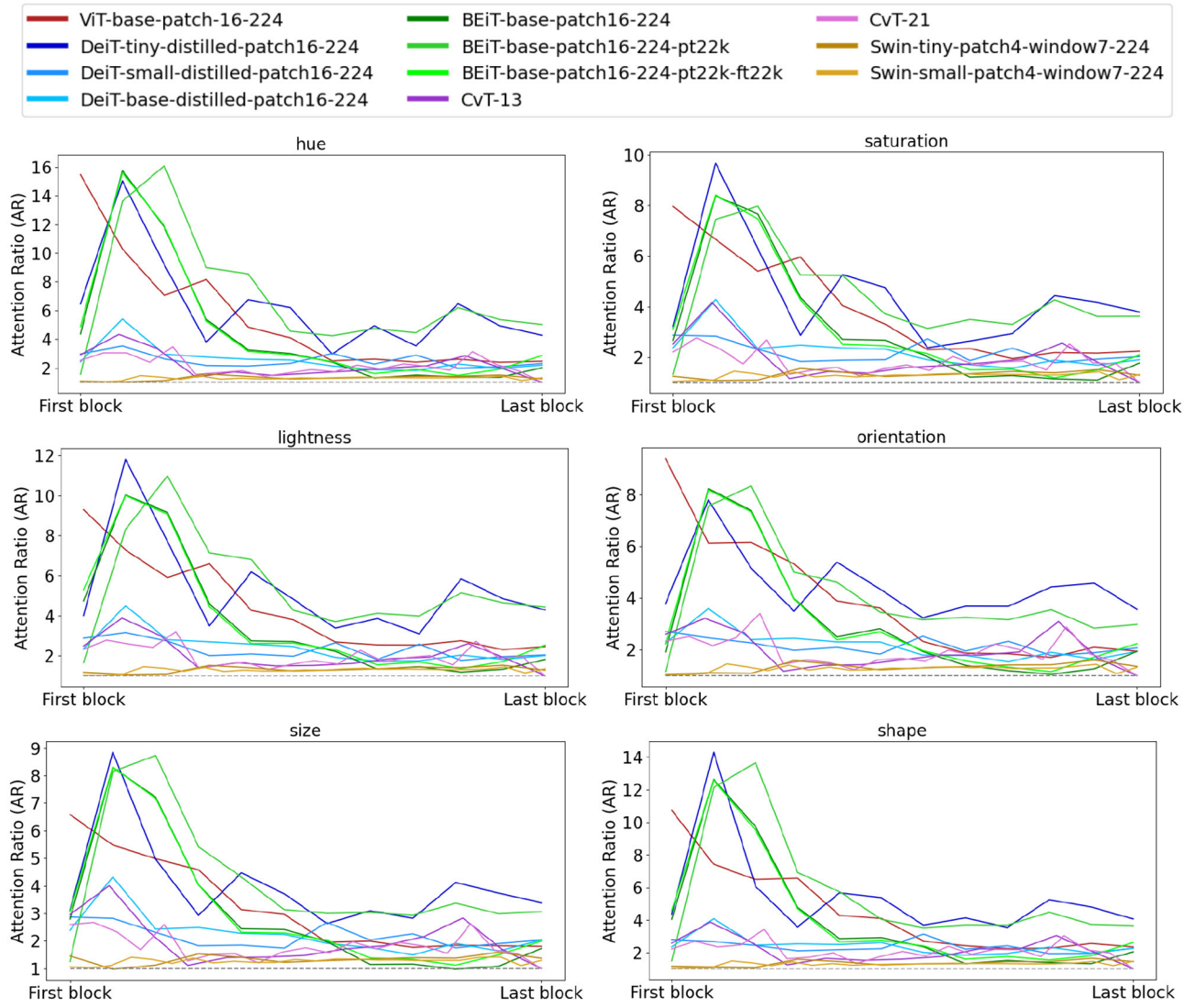


FIGURE 6
Mean attention ratio for all the architectures from Table 1 plotted separately for each visual feature that defined the perceptual grouping of figures in the input. Similar to Figure 5, and for ease of comparison, we plotted the *AR* for all models within the same range on the *x*-axis. Interestingly, Swin and CvT, two models that adapted ViT to the visual domain, have relatively smaller attention ratios compared to the rest of the architectures, suggesting that incorporating scale and shifting the token position in Swin and convolution in CvT architectures results in mixing of figure and background representations and consequently attention scores. Among the other architectures that use ViT as the underlying model, the attention ratio plots are somewhat similar to those of Figure 4B, that is, larger attention ratios in earlier blocks with a decline in the hierarchy.

relative to attention-based maps, in comparison with convolutional saliency models (see Kotseruba et al., 2019, their Figure 3), those performed relatively similar to convolution-based models. Together, these results suggest that contrary to the expectation, the proposed attention mechanisms in vision transformers are not advantageous vs. convolutional computations in representing visual salience.

3.2.2. The O^3 dataset results

We measured the maximum saliency ratios MSR_{targ} and MSR_{bg} for feature and attention-based saliency maps of all blocks of vision transformers in Table 1. These ratios are plotted in Figure 11, demonstrating poor performance of all models in detecting the target in natural images of the O^3 dataset. We acknowledge that we expected improved performance of vision transformers on the O^3 dataset with natural images compared to the results on synthetic stimuli of the P^3 dataset. However, whereas MSR_{targ} ratios larger than 1 are expected (higher salience of target vs. distractors), in both feature and attention-based saliency maps, the ratios were distinctly below 1 across all blocks of all models, with the exception of later blocks of two BEiT architectures. Notable are the feature-based ratios of ViT-base-patch16-224 with peaks in earlier blocks and a steep decrease in higher layers. In contrast, all three BEiT architectures show the opposite behavior and perform poorly in earlier blocks but correct the ratio in mid-higher stages of processing.

The MSR_{bg} ratios illustrated in Figure 11 follow a similar theme as MSR_{targ} ratios. Even though MSR_{bg} ratios <1 suggest that the target is deemed more salient than the background, most of these

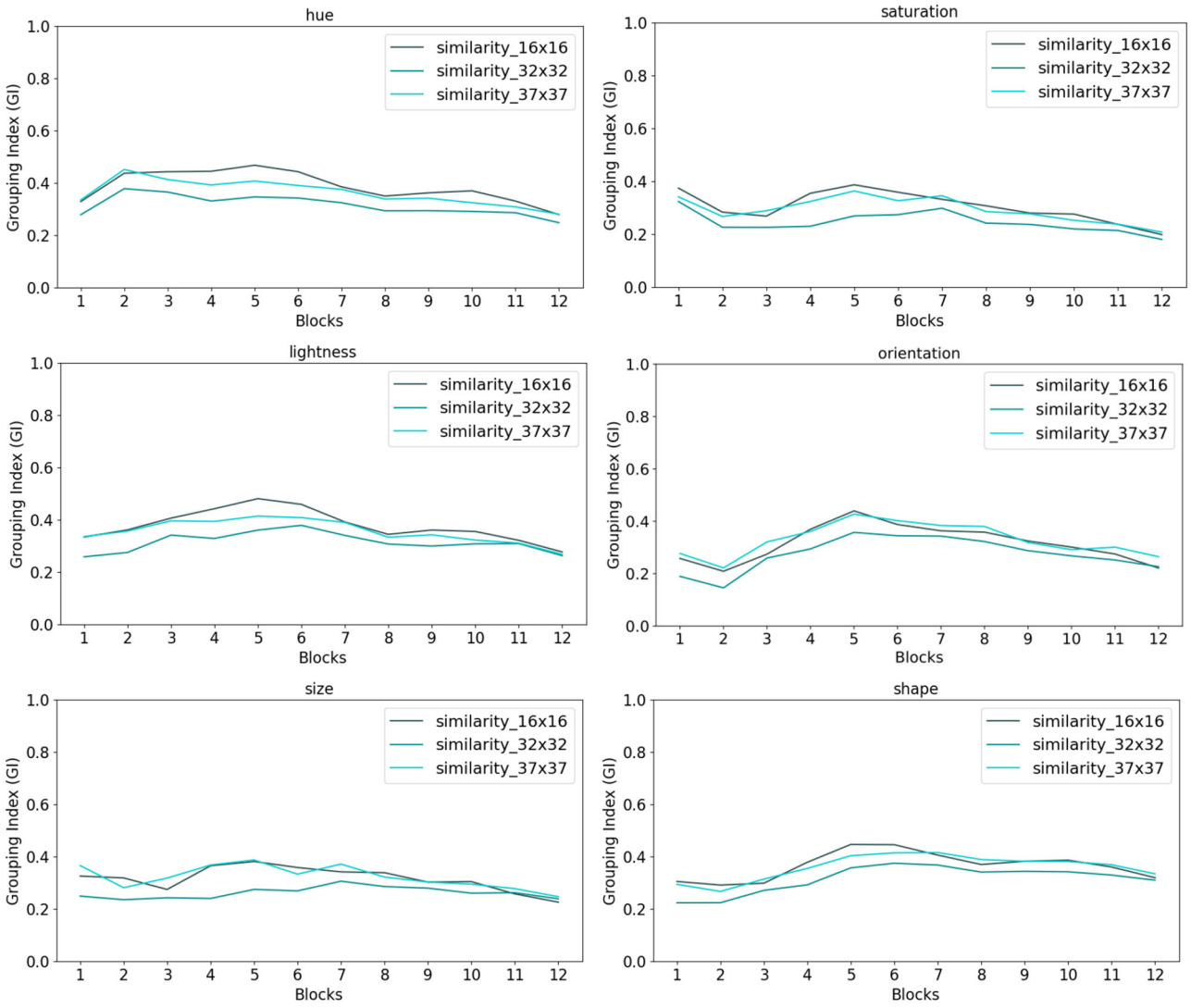


FIGURE 7
Each token in ViT-based architectures has a fixed position and size across the hierarchy of transformer encoders. This property is noted as a shortcoming of some vision transformers. To control for position and size of tokens in these models, we designed our grouping dataset according to the ViT model tokens such that each figure in the stimulus would fit within and positioned inside the model 16 × 16 tokens. To test for the effect of figure size, we introduced a version of grouping dataset with figures centered at every other ViT token but larger in size such that each figure would fit a 32 × 32 square. We also introduced a third version where figures in the stimuli were token-agnostic. In the third version, the set of figures occupy the center of image and each figure fits within a 37 × 37 square. We tested the grouping performance of the DeiT-base-distilled-patch16-224 architecture over all three versions of the dataset. Note that DeiT-base-distilled-patch16-224 utilizes an identical architecture as ViT-base-patch16-224 with a different training regime. Our results over the various visual features in the dataset demonstrate comparable results over the three versions of the dataset, suggesting no significant effect of token position or size in grouping in vision transformers.

models have MSR_{bg} ratios larger than 1 in their hierarchy. Among all models, feature-based saliency of BEiT and Swin architectures have the best overall performance.

For a few randomly selected images from the O^3 dataset, Figures 12–14 demonstrate the attention-based saliency map of the block with best MSR_{targ} ratio for each model. Each saliency map in these figures is scaled to the original image size for demonstration purposes. Interestingly, saliency maps in Figure 13 show how the same BEiT model with varying training result in vastly different attention-based maps.

To summarize, for a bottom-up architecture that is claimed to implement attention mechanisms, we expected a boost in saliency detection compared to convolution-based models with no explicit attention modules. Our results on the O^3 dataset, however, point to the contrary, specifically in comparison with the best ratios reported in Kotseruba et al. (2019) for MSR_{targ} and MSR_{bg} at 1.4 and 1.52, respectively. These results, together with the proposal of Liu et al. (2022) for a modernized convolution-based model with comparable performance to vision transformers, overshadow the claim of attention mechanisms in these models.

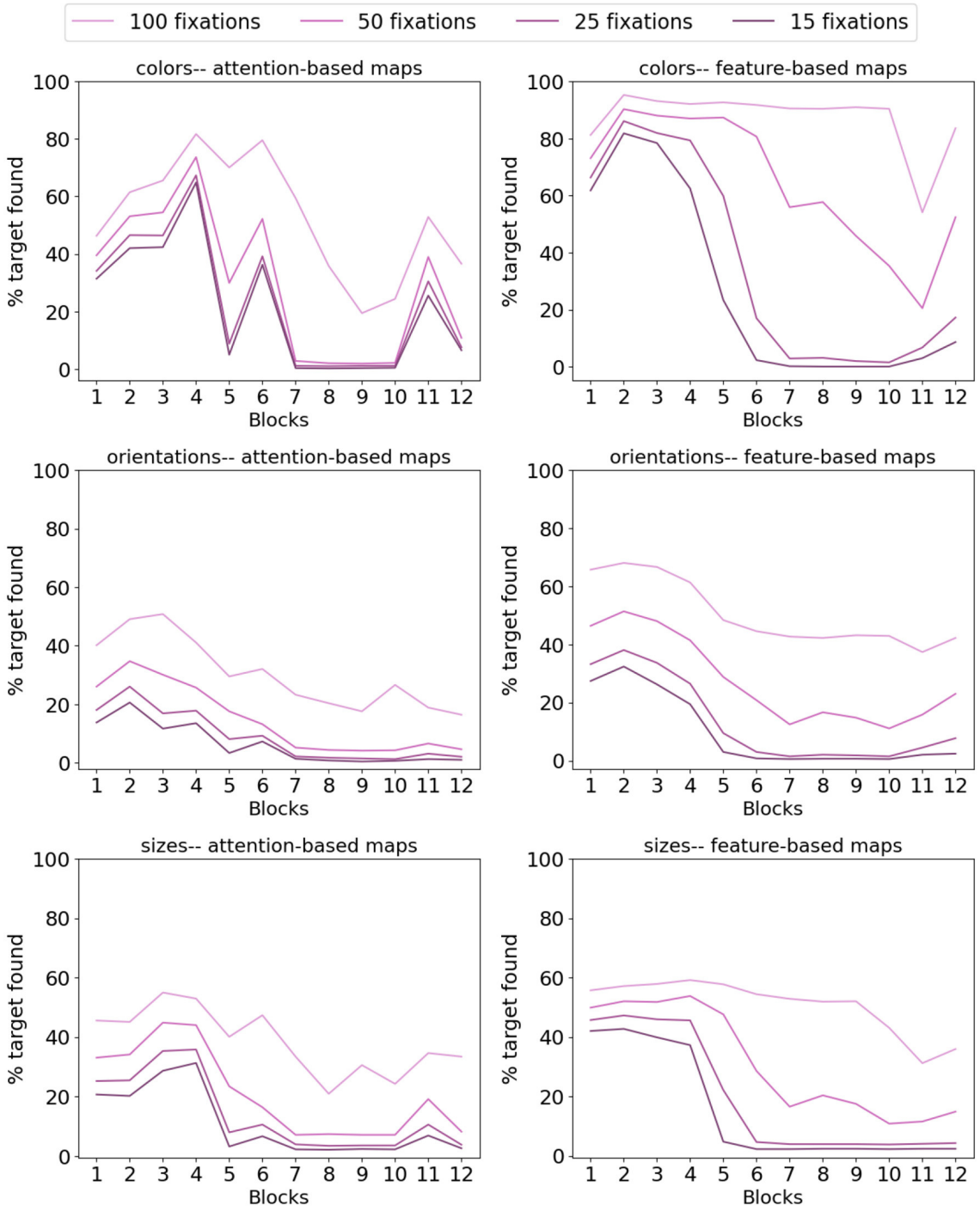


FIGURE 8
Target detection rate of the ViT-base-patch16-224 model for 15, 25, 50, and 100 fixations on images of the P^3 dataset. Legend on top applies to all plots. For this model with 16 × 16 pixels tokens, each masking after a fixation masks almost an entire token. Therefore, chance performance will be at 6, 10, 20, and 40% for 15, 25, 50, and 100 fixations. Comparing the plots of the left column for attention-based saliency maps vs. those on the right obtained from feature-based saliency maps indicates superior performance of feature-based maps for salient target detection. This is interesting in that modules claimed to implement attention mechanisms are expected to succeed in detecting visually salient figures in the input. Overall, for both attention and feature-based maps, color targets have higher detection rates vs. orientation and size, the conditions in which performance is mainly at chance level for all fixation thresholds and across all blocks in the ViT hierarchy. Additionally, in both attention and feature-based maps, performance peaks in earlier blocks and declines in later layers, suggesting multiple transformer encoder blocks mix representations across spatial locations such that the model cannot detect the visually salient target almost immediately or even by chance.

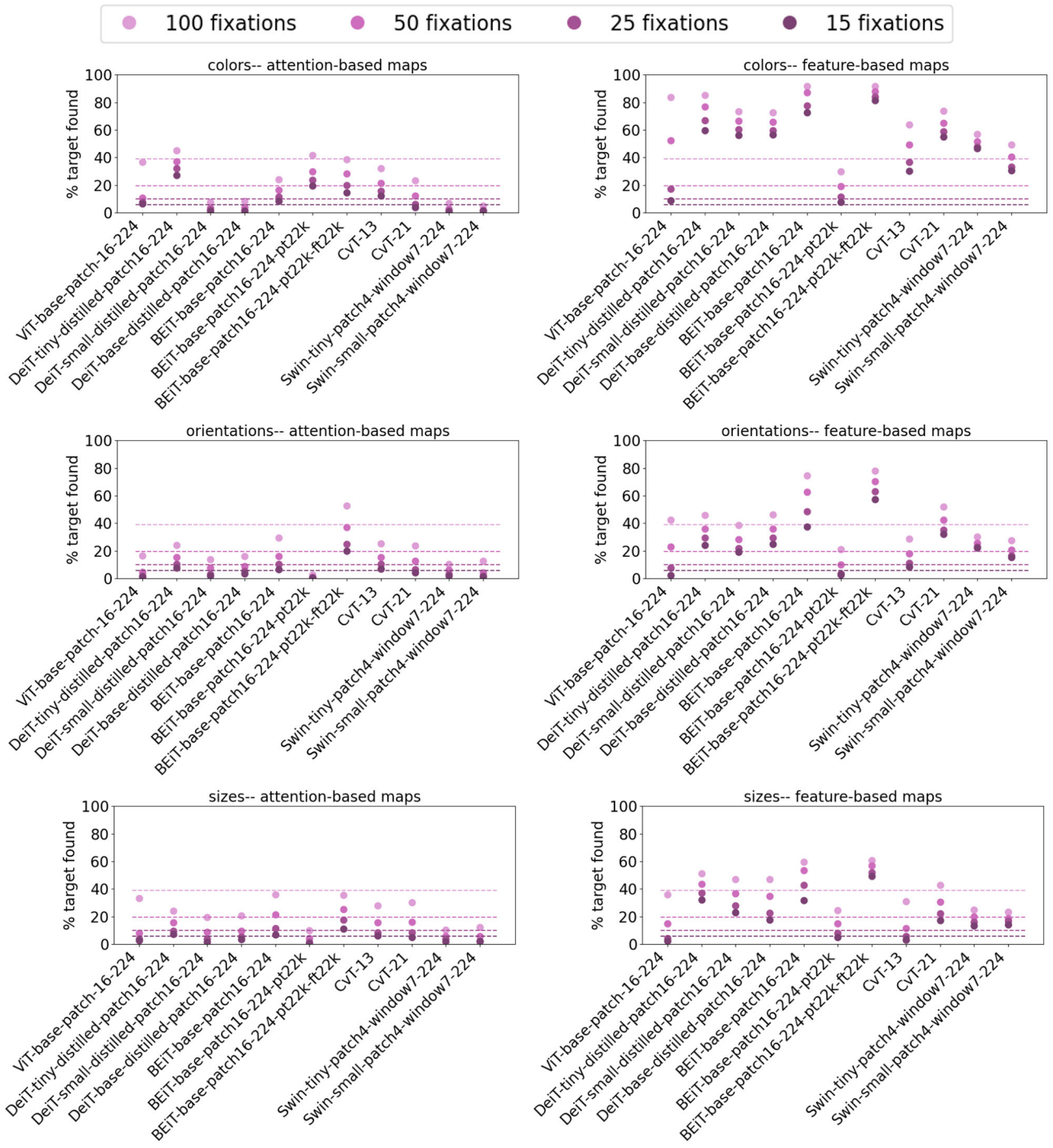


FIGURE 9
Target detection rate of the last block of vision transformers investigated in this work for 15, 25, 50, and 100 fixations on stimuli from the P^3 dataset. The chance level performance of the ViT model is plotted as dashed lines with matching colors for each fixation threshold. Similar to the observation of ViT, feature-based maps outperform attention-based maps and generally at rates higher than chance. Color targets are easier to detect for both map types. Interestingly, both Swin architectures struggle to detect size targets in both attention and feature-based maps, despite incorporating multiple scales in their model.

4. Discussion

Our goal in this work was to investigate if the self-attention modules in vision transformers have similar effects to human attentive visual processing. Vision transformers have attracted much interest in the past few years partly due to out-performing CNNs in various visual tasks, and in part due to incorporating modules that were claimed to implement attention mechanisms. Specifically, the origins of attention mechanisms in transformers could be traced back to the work by Xu et al. (2015), where they

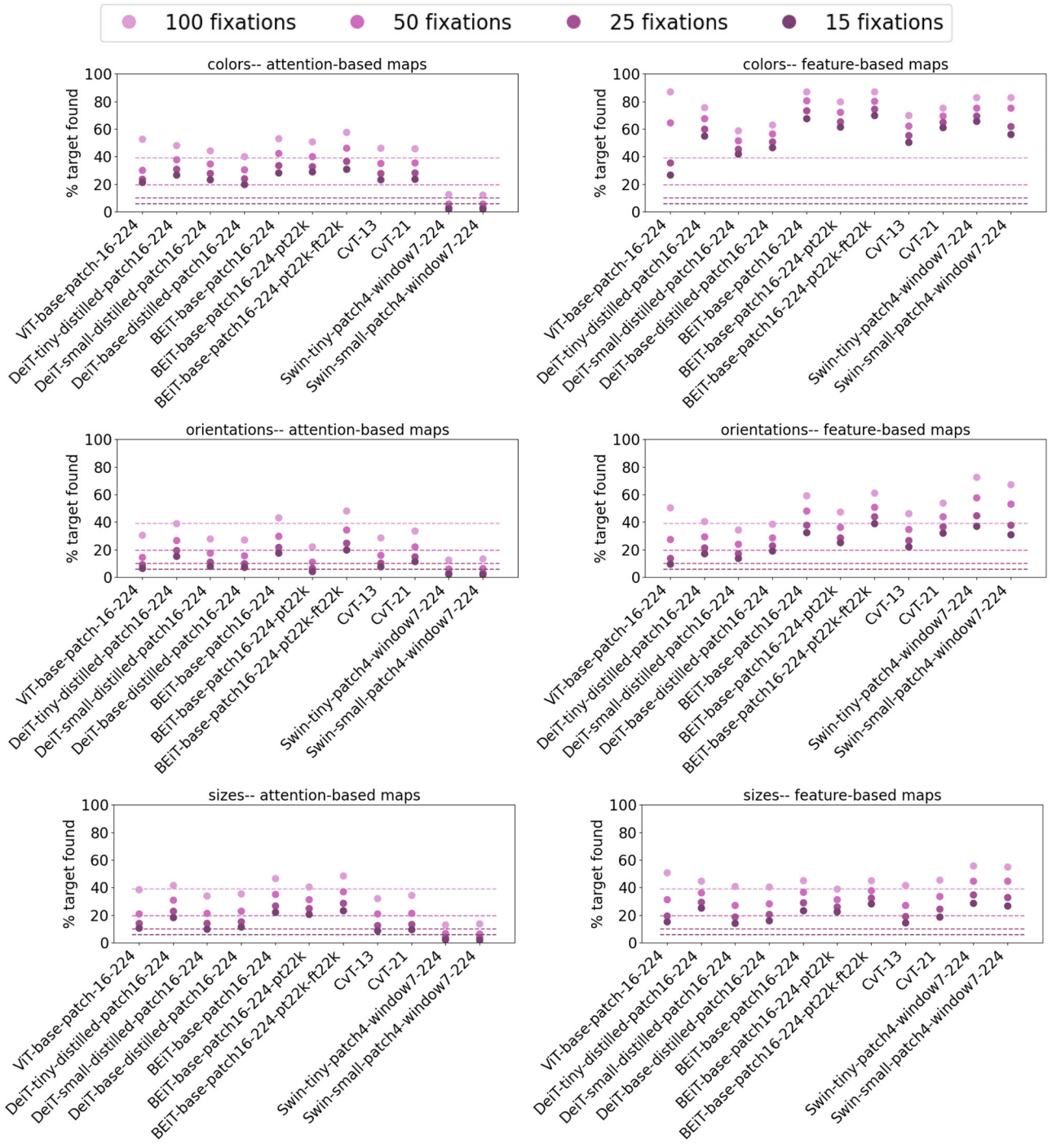


FIGURE 10
Average target detection rate over all blocks of vision transformers for 15, 25, 50, and 100 fixations on stimuli from the P^3 dataset. Compared to detection rates in Figure 9, mean detection rates are higher for all models in all conditions (color, size, and orientation), indicating superior performance of earlier transformer blocks compare to the final block in these models. In line with results in Figures 8, 9, color targets are easier to detect and that generally, feature-based maps outperform attention-based maps in salient target detection.

introduced an attention-based model for image captioning. Xu et al. (2015) motivated modeling attention in their network by reference to attention in the human visual system and its effect that "allows for salient features to dynamically come to the forefront as needed", especially in the presence of clutter in the input. In light of these observations, a curious question to ask is if these computational attention mechanisms have similar effects as their source of inspiration. Despite some previous attempts (Naseer et al., 2021; Tuli et al., 2021; Park and Kim, 2022), the role and effect of the attention modules in vision

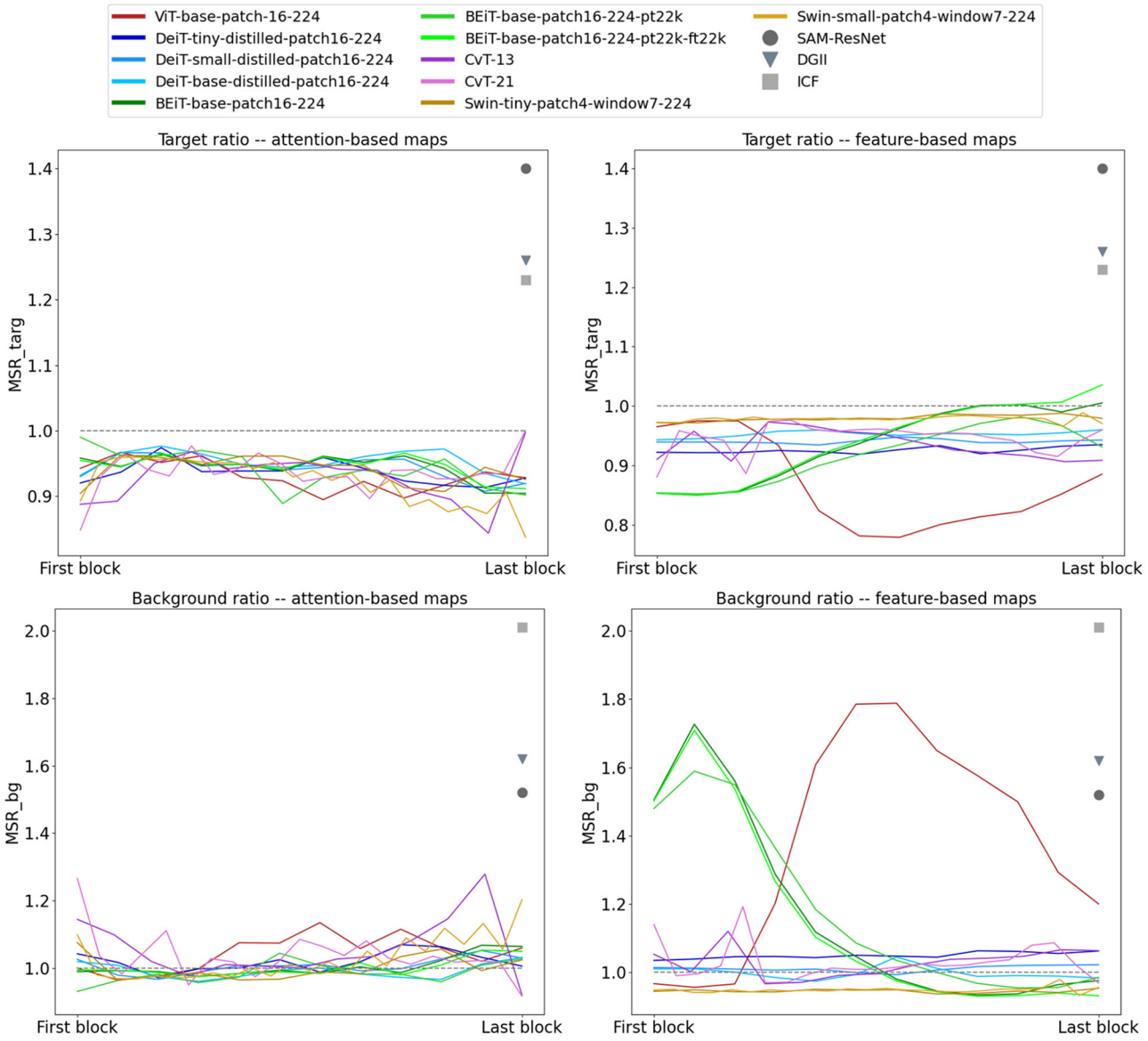


FIGURE 11
Target and background saliency ratios (MSR_{targ} and MSR_{bg}) for all vision transformer architectures on natural images of the O^3 dataset. Even though these models were trained on natural images, they fail in assigning higher saliency to the target object vs. the distractors or background (smaller than 1 MSR_{targ} and larger than 1 MSR_{bg} ratios) in both feature and attention-based maps. For easier comparison, we also plotted MSR_{targ} and MSR_{bg} for the top-3 best performing deep saliency models from Kotseruba et al. (2019). These results compared to the reported ratios in Kotseruba et al. (2019) for traditional and deep convolution-based saliency models suggest that the proposed attention mechanisms do not enhance the performance of vision transformers when the goal is to detect the visually salient object in the stimulus.

transformers have been largely unknown. To give a few examples, in a recent work, Li et al. (2023) studied the interactions of the attention heads and the learned representations in multi-head attention modules and reported segregation of representations across heads. (Abnar and Zuidema, 2020) investigated the effect of various approaches for visualizing attention map as an interpretability step and with their attention rollout approach often employed for this purpose. Ghiasi et al. (2022) visualized the learned representations in vision transformers and found similarity to those of CNNs. In contrast, Caron et al. (2021) and Raghu et al. (2021) reported dissimilarities in learned representations across the hierarchy of vision transformers and CNNs. Cordonnier et al. (2020) as well as some others (D'Ascoli et al., 2021) suggested attention mechanisms as a generalized form of convolution. The quest to understand the role and effect of attention modules in transformers is still ongoing as these models are relatively new and the notable variations in findings (for example, dis/similarity to CNNs) adds to its importance. Yet, and to the best of our knowledge, none of these studies investigated if the computations in self-attention modules would have similar effects on visual processing as those discovered with visual attention in humans.

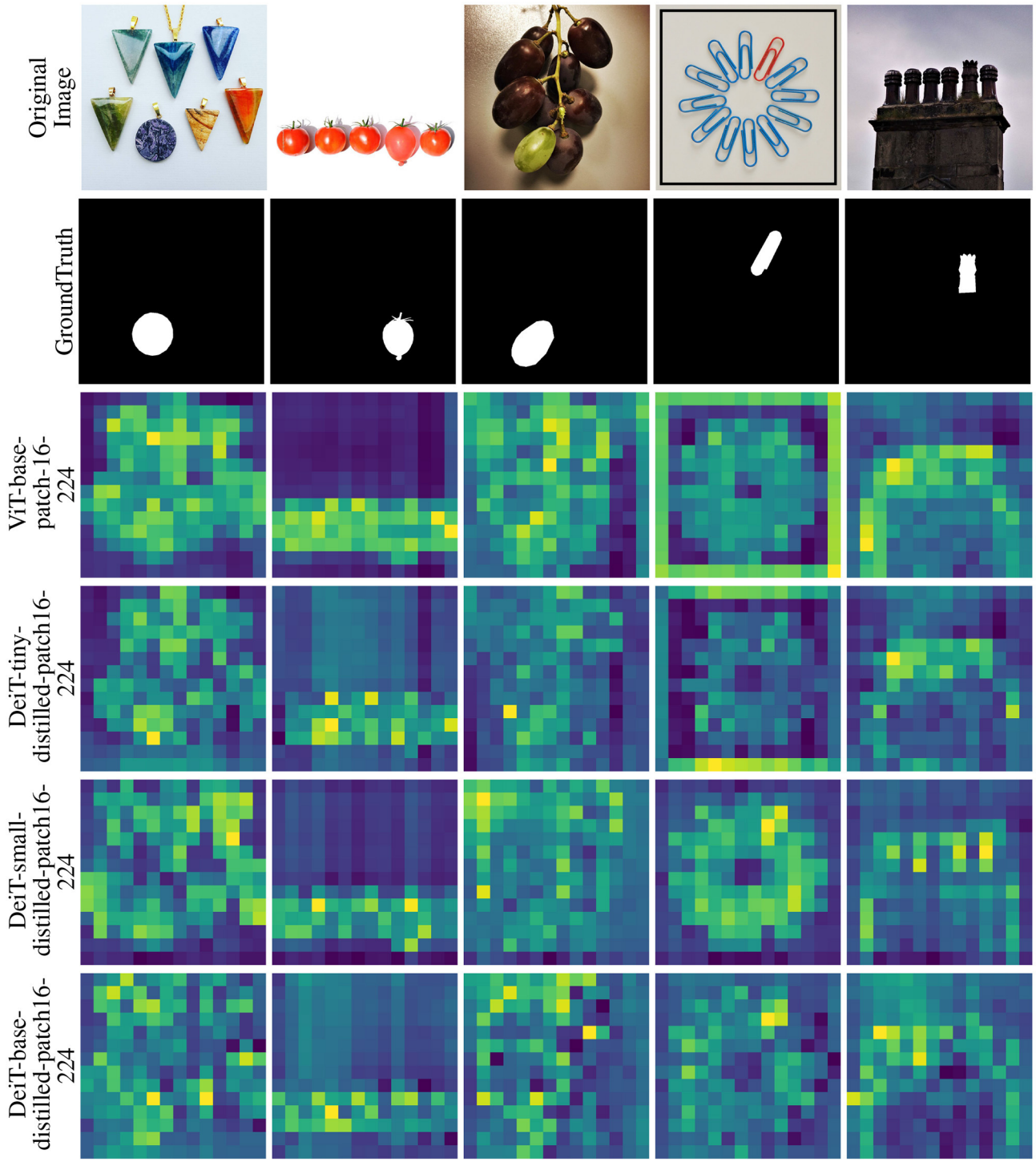


FIGURE 12
The attention-based saliency map of the block with the best MSR_{targ} ratio for ViT and DeiT models on a select few images from the O^3 dataset. In each map, saliency varies from blue (least salient) to yellow (most salient). Each saliency map in these figures is scaled to the original image size for demonstration purposes.

In this work, we studied two aspects of processing in vision transformers: the formulation of attention in self-attention modules, and the overall bottom-up architecture of these deep neural architectures. Our investigation of attention formulation in vision transformers suggested that these modules perform Gestalt-like similarity grouping in the form of horizontal relaxation labeling whereby interactions from multiple spatial positions determine the update in the representation of a token. Additionally,

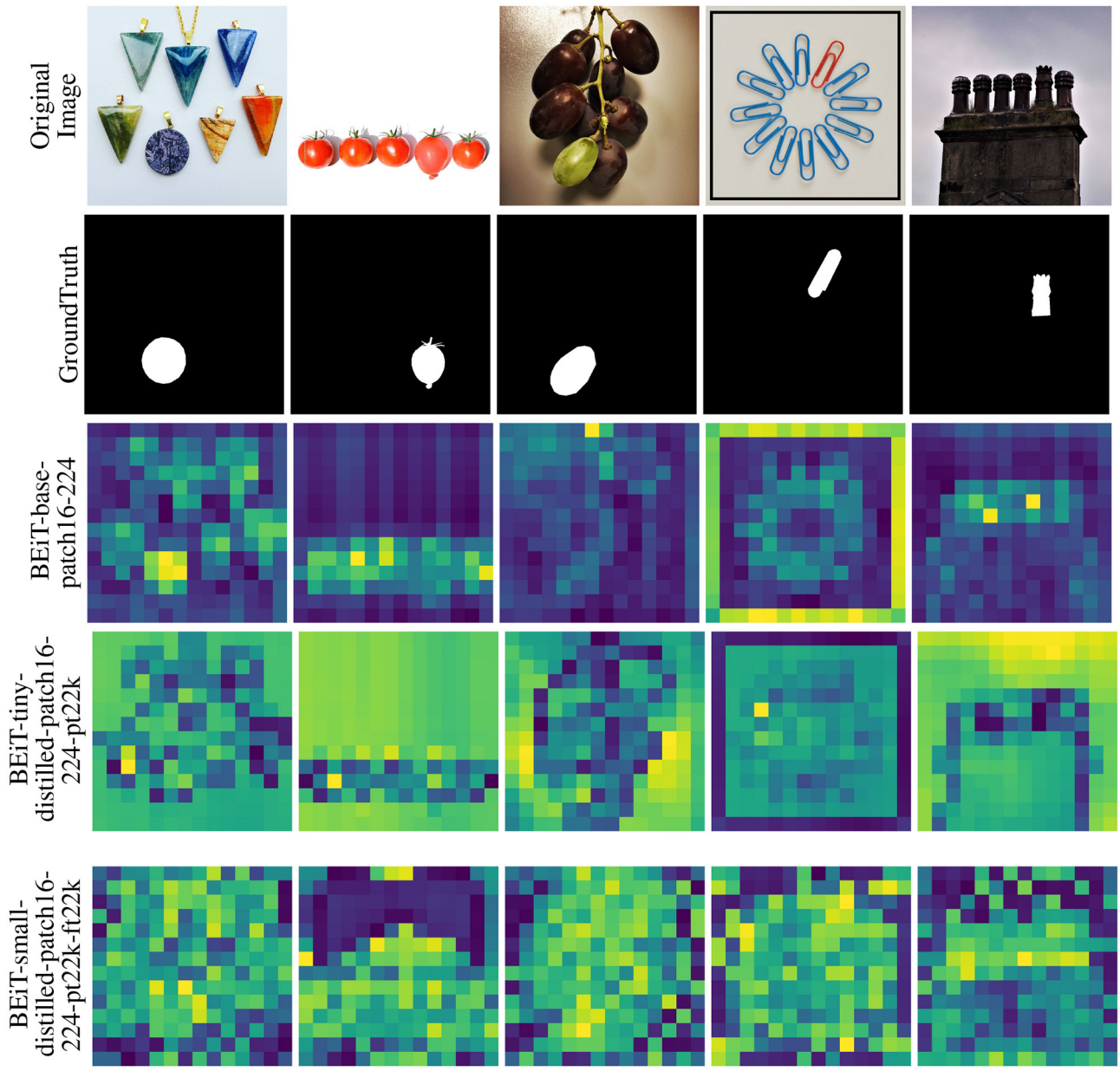


FIGURE 13
The attention-based saliency map of the block with the best MSR_{targ} ratio for BEiT models on a select few images from the O^3 dataset. In each map, saliency varies from blue (least salient) to yellow (most salient). Each saliency map in these figures is scaled to the original image size for demonstration purposes. An interesting observation is how the variants of the same model with differing training regimes result in vastly different attention-based saliency maps.

given previous evidence on the role of feedback in human visual attention (Folk et al., 1992; Bacon and Egeth, 1994; Desimone and Duncan, 1995; Kim and Cave, 1999; Yantis and Egeth, 1999; Lamy et al., 2003; Connor et al., 2004; Baluch and Itti, 2011; Peterson, 2015), we argued that if vision transformers implement attention mechanisms, those can only be in the form of bottom-up and stimulus-driven visual salience signals.

Testing a family of vision transformers on a similarity grouping dataset suggested that the attention modules in these architectures perform similarity grouping and that the effect decays as hierarchical level increases in the hierarchy especially because more non-figure tokens are grouped with figures in the stimulus over multiple transformer encoder blocks. Most surprising, however, were our findings in the task of singleton detection as a canonical example of saliency detection. With both synthetic and natural stimuli, vision transformers demonstrated sub-optimal performance in comparison with traditional and deep convolution-based saliency models.

The P^3O^3 dataset was designed according to psychological and neuroscience findings on human visual attention.

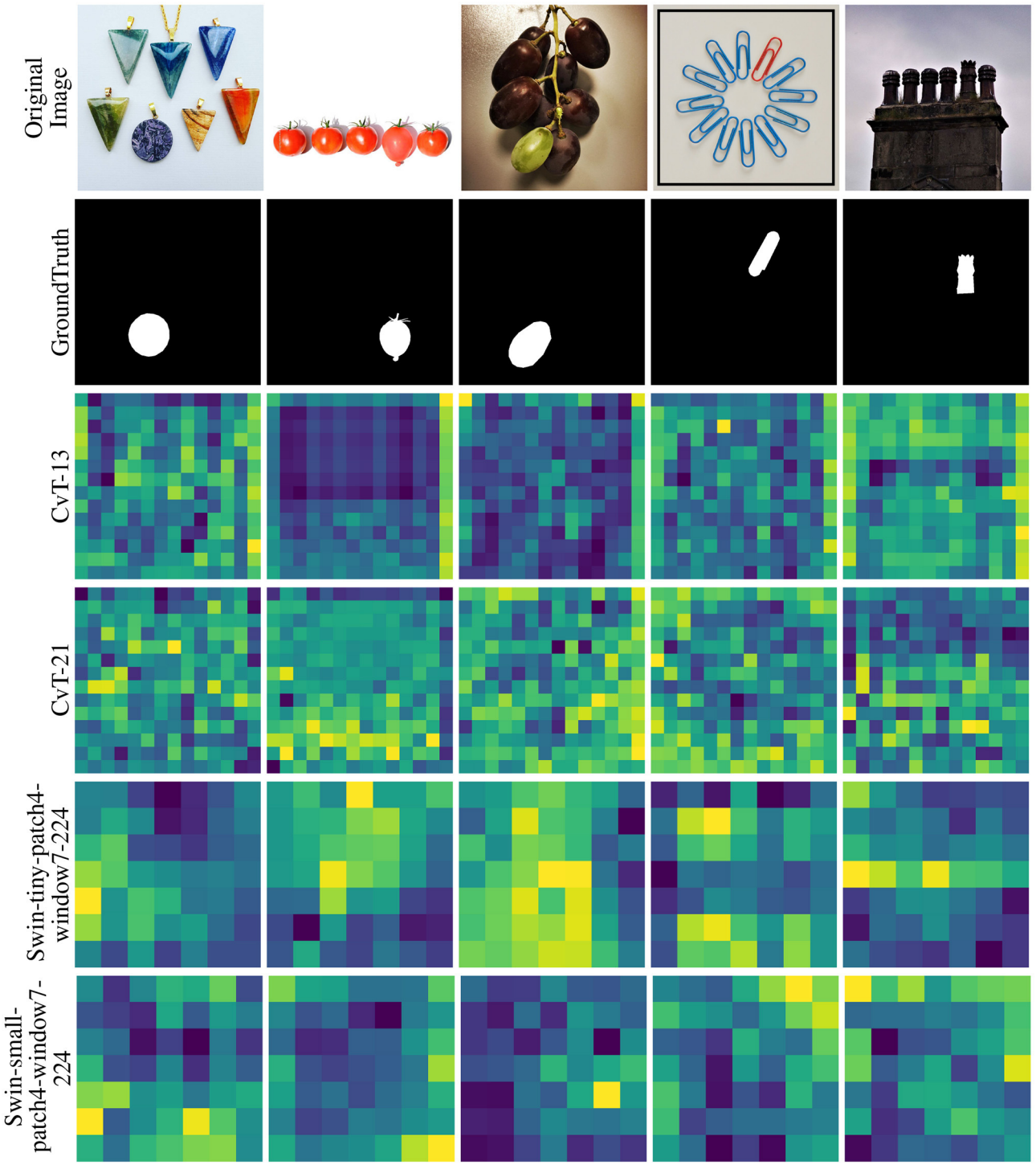


FIGURE 14
The attention-based saliency map of the block with the best MSR_{targ} ratio for CvT and Swin models on a select few images from the O^3 dataset. In each map, saliency varies from blue (least salient) to yellow (most salient). Each saliency map in these figures is scaled to the original image size for demonstration purposes.

Kotseruba et al. (2019) demonstrated a gap between human performance and traditional/CNN-based saliency models in singleton detection tasks. The fact that Kotseruba et al. (2019) reported that training CNN-based saliency models on these stimuli did not improve their performance, hints on a more fundamental difference between the two systems. Several other

works have provided evidence on the lack of human equivalence in deep neural networks (Ghodrati et al., 2014; Dodge and Karam, 2017; Kim et al., 2018; Geirhos et al., 2019; Horikawa et al., 2019; Hu et al., 2019; RichardWebster et al., 2019; Wloka and Tsotsos, 2019; Baker et al., 2020; Lonnqvist et al., 2021; Ricci et al., 2021; Xu and Vaziri-Pashkam, 2021a,b, 2022; Ayzenberg and Lourenco, 2022; Feather et al., 2022; Fel et al., 2022; Vaishnav et al., 2022; Zerroug et al., 2022; Zhou Q. et al., 2022) on various aspects of visual processing. The claim of implementing attention mechanisms in vision transformers offered the possibility that these models might be more human-like. This impression was confirmed in the work of Tuli et al. (2021) who reported that vision transformers are more human-like than CNNs based on performance on the Stylized ImageNet dataset (Geirhos et al., 2019). Our work, however, adds to the former collection of studies and reveals a gap between human visual attention and the mechanisms implemented in vision transformers.

This work can be further extended in several directions. For example, even though Kotseruba et al. (2019) found training CNN-based saliency models on the O^3 dataset did not improve their saliency detection performance, an interesting experiment is to fine-tune vision transformers on the O^3 dataset and evaluate the change or lack of change in their saliency detection performance. Additionally, incorporating vertical visual processes into the formulation in Equation (1) is another avenue to explore in the future.

To conclude, not only does our deliberate study of attention formulation and the underlying architecture of vision transformers suggest that these models perform perceptual grouping and do not implement attention mechanisms, but also our experimental evidence, especially from the P^3O^3 datasets confirms those observations. The mechanisms implemented in self-attention modules of vision transformers can be interpreted as *lateral interactions* within a single layer. In some architectures, such as ViT, the entire input defines the neighborhood for these lateral interactions, in some others (Yang et al., 2021) this neighborhood is limited to local regions of input. Although Liu et al. (2022) found similar performance in a modernized CNNs, the ubiquity of lateral interactions in the human and non-human primate visual cortex (Stettler et al., 2002; Shushruth et al., 2013) suggest the importance of these mechanisms in visual processing. Our observation calls for future studies to investigate whether vision transformers show the effects that are commonly attributed to lateral interactions in the visual cortex such as crowding, tilt illusion, perceptual filling-in, etc. (Lin et al., 2022). Self-attention in vision transformers performs perceptual organization using feature similarity grouping, not attention. Additionally, considering Gestalt principles of grouping, vision transformers implement a narrow aspect of perceptual grouping, namely similarity, and other aspects such as symmetry and proximity seem problematic for these models. The term attention has a long history going back to the 1800's and earlier (see Berlyne, 1974) and in computer vision to 1970's (for examples, see Hanson and Riseman, 1978). With decades of research on biological and computational aspects of attention, the confusion caused by inappropriate use of terminology and technical term conflation has already been problematic. Therefore, we remain with the suggestion that even though vision transformers do not perform attention as claimed, they incorporate visual mechanisms in deep architectures that were previously absent in CNNs and provide new opportunities for further improvement of our computational vision models.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

PM and JT developed the theoretical formalisms, analyzed the data, and wrote the manuscript. PM contributed to the implementation, designed, and carried out the experiments. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by several sources for which the authors are grateful: Air Force Office of Scientific Research [grant numbers FA9550-18-1-0054 and FA9550-22-1-0538 (Computational Cognition and Machine Intelligence, and Cognitive and Computational Neuroscience Portfolios)], the Canada Research Chairs Program (grant number 950-231659), and the Natural Sciences and Engineering Research Council of Canada (grant numbers RGPIN-2016-05352 and RGPIN-2022-04606).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2023.1178450/full#supplementary-material

References

Abnar, S., and Zuidema, W. (2020). "Quantifying attention flow in transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4190–4197.

Anderson, B. (2023). Stop paying attention to "attention". *Wiley Interdiscip. Rev. Cogn. Sci*. 14, e1574. doi: 10.1002/wcs.1574

Ayzenberg, V., and Lourenco, S. (2022). Perception of an object's global shape is best described by a model of skeletal structure in human infants. *Elife*. 11, e74943. doi: 10.7554/eLife.74943

Bacon, W. F., and Egeth, H. E. (1994). Overriding stimulus-driven attentional capture. *Percept. Psychophys*. 55, 485–496. doi: 10.3758/BF03205306

Baker, N., Lu, H., Erlikhman, G., and Kellman, P. J. (2020). Local features and global shape information in object classification by deep convolutional neural networks. *Vision Res*. 172, 46–61. doi: 10.1016/j.visres.2020.04.003

Baluch, F., and Itti, L. (2011). Mechanisms of top-down attention. *Trends Neurosci*. 34, 210–224. doi: 10.1016/j.tins.2011.02.003

Bao, H., Dong, L., Piao, S., and Wei, F. (2022). "BEit: BERT pre-training of image transformers," in *International Conference on Learning Representations*.

Berlyne, D. E. (1974). "Attention," in *Handbook of Perception*., *Chapter 8*, eds E. C. Carterette, and M. P. Friedman (New York, NY: Academic Press).

Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., and Veit, A. (2021). "Understanding robustness of transformers for image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10231–10241.

Borji, A., and Itti, L. (2012). State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell*. 35, 185–207. doi: 10.1109/TPAMI.2012.89

Bruce, N. D., Wloka, C., Frosst, N., Rahman, S., and Tsotsos, J. K. (2015). On computational modeling of visual saliency: examining what's right, and what's left. *Vision Res*. 116, 95–112. doi: 10.1016/j.visres.2015.01.010

Bylinskii, Z., DeGennaro, E. M., Rajalingham, R., Ruda, H., Zhang, J., and Tsotsos, J. K. (2015). Towards the quantitative evaluation of visual attention models. *Vision Res*. 116:258–268. doi: 10.1016/j.visres.2015.04.007

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., et al. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Comput. Biol*. 10, e1003963. doi: 10.1371/journal.pcbi.1003963

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., et al. (2021). "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 9650–9660.

Carrasco, M. (2011). Visual attention: the past 25 years. *Vision Res*. 51, 1484–1525. doi: 10.1016/j.visres.2011.04.012

Chen, C.-F. R., Fan, Q., and Panda, R. (2021). "Crossvit: cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 357–366.

Connor, C. E., Egeth, H. E., and Yantis, S. (2004). Visual attention: bottom-up versus top-down. *Curr. Biol*. 14, R850–R852. doi: 10.1016/j.cub.2004.09.041

Cordonnier, J.-B., Loukas, A., and Jaggi, M. (2020). "On the relationship between self-attention and convolutional layers," in *Eighth International Conference on Learning Representations-ICLR 2020, number CONF*.

Dai, Z., Liu, H., Le, Q. V., and Tan, M. (2021). "CoAtNet: marrying convolution and attention for all data sizes," in *Advances in Neural Information Processing Systems, Vol. 34*, eds M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan (Curran Associates, Inc.), 3965–3977.

D'Ascoli, S., Touvron, H., Leavitt, M. L., Morcos, A. S., Biroli, G., and Sagun, L. (2021). "ConViT: improving vision transformers with soft convolutional inductive biases," in *Proceedings of the 38th International Conference on Machine Learning, Vol. 139*, eds M. Meila, and T. Zhang (PMLR), 2286–2296.

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Ann. Rev. Neurosci*. 18, 193–222. doi: 10.1146/annurev.ne.18.030195.001205

Di Lollo, V. (2018). Attention is a sterile concept; iterative reentry is a fertile substitute. *Conscious. Cogn*. 64:45–49. doi: 10.1016/j.concog.2018.02.005

Dodge, S., and Karam, L. (2017). "A study and comparison of human and deep learning recognition performance under visual distortions," in *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, 1–7.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). "An image is worth 16x16 words: transformers for image recognition at scale," in *International Conference on Learning Representations*.

Eickenberg, M., Gramfort, A., Varoquaux, G., and Thirion, B. (2017). Seeing it all: convolutional network layers map the function of the human visual system. *Neuroimage*. 152, 184–194. doi: 10.1016/j.neuroimage.2016.10.001

Feather, J., Leclerc, G., Mądry, A., and McDermott, J. H. (2022). Model metamers illuminate divergences between biological and artificial neural networks. *bioRxiv, pages* 2022–05. doi: 10.32470/CCN.2022.1147-0

Fel, T., Rodriguez, I. F. R., Linsley, D., and Serre, T. (2022). "Harmonizing the object recognition strategies of deep neural networks with humans," in *Advances in Neural Information Processing Systems*, eds A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, K.

Folk, C. L., Remington, R. W., and Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *J. Exp. Psychol. Hum. Percept. Perform*. 18, 1030. doi: 10.1037/0096-1523.18.4.1030

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2019)."Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in *International Conference on Learning Representations*.

Ghiasi, A., Kazemi, H., Borgnia, E., Reich, S., Shu, M., Goldbulm, A., et al. (2022). What do vision transformers learn? A visual exploration. arXiv [Preprint]. *arXiv: 2212.06727*. Available online at: https://arxiv.org/pdf/2212.06727.pdf

Ghodrati, M., Farzmahdi, A., Rajaei, K., Ebrahimpour, R., and Khaligh-Razavi, S.-M. (2014). Feedforward object-vision models only tolerate small image variations compared to human. *Front. Comput. Neurosci*. 8, 74. doi: 10.3389/fncom.2014.00074

Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., et al. (2022)."Cmt: convolutional neural networks meet vision transformers," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12165–12175.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., et al. (2022). "A survey on vision transformer," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45 (IEEE), 87–110.

Hanson, A., and Riseman, E. (1978). *Computer Vision Systems: Papers from the Workshop on Computer Vision Systems*. Amherst, MA: Held at the University of Massachusetts, Academic Press.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016* (Las Vegas, NV: IEEE Computer Society), 770–778.

Herzog, M. H., and Clarke, A. M. (2014). Why vision is not both hierarchical and feedforward. *Front. Comput. Neurosci*. 8, 135. doi: 10.3389/fncom.2014.00135

Hommel, B., Chapman, C. S., Cisek, P., Neyedli, H. F., Song, J.-H., and Welsh, T. N. (2019). No one knows what attention is. *Attent. Percept. Psychophys*. 81, 288–2303. doi: 10.3758/s13414-019-01846-w

Horikawa, T., Aoki, S. C., Tsukamoto, M., and Kamitani, Y. (2019). Characterization of deep neural network features by decodability from human brain activity. *Sci. Data* 6, 1–12. doi: 10.1038/sdata.2019.12

Hu, B., Khan, S., Niebur, E., and Tripp, B. (2019). "Figure-ground representation in deep neural networks," in *2019 53rd Annual Conference on Information Sciences and Systems (CISS)* (IEEE), 1–6.

Itti, L. (2007). Visual salience. *Scholarpedia*. 2, 3327. doi: 10.4249/scholarpedia.3327

Itti, L., Rees, G., and Tsotsos, J. K. (2005). *Neurobiology of Attention*. Elsevier.

Kastner, S., and Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annu. Rev. Neurosci*. 23, 315–341. doi: 10.1146/annurev.neuro.23.1.315

Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput. Biol*. 10, e1003915. doi: 10.1371/journal.pcbi.1003915

Kim, J., Ricci, M., and Serre, T. (2018). Not-so-clevr: learning same-different relations strains feedforward neural networks. *Interface Focus* 8, 20180011. doi: 10.1098/rsfs.2018.0011

Kim, M.-S., and Cave, K. R. (1999). Top-down and bottom-up attentional control: on the nature of interference from a salient distractor. *Percept. Psychophys*. 61, 1009–1023. doi: 10.3758/BF03207609

Kim, Y., Denton, C., Hoang, L., and Rush, A. M. (2017). "Structured attention networks," in *International Conference on Learning Representations*.

Knudsen, E. I. (2007). Fundamental components of attention. *Annu. Rev. Neurosci*. 30, 57–78. doi: 10.1146/annurev.neuro.30.051606.094256

Kotseruba, I., Wloka, C., Rasouli, A., and Tsotsos, J. K. (2019). "Do saliency models detect odd-one-out targets? New datasets and evaluations," in *British Machine Vision Conference (BMVC)*.

Krauzlis, R. J., Wang, L., Yu, G., and Katz, L. N. (2023). What is attention? *Wiley Interdiscip. Rev. Cogn. Sci*. 14, e1570. doi: 10.1002/wcs.1570

Kubilius, J., Bracci, S., and Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput. Biol*. 12, e1004896. doi: 10.1371/journal.pcbi.1004896

Lamy, D., Tsal, Y., and Egeth, H. E. (2003). Does a salient distractor capture attention early in processing? *Psychonom. Bull. Rev*. 10, 621–629. doi: 10.3758/BF03196524

Li, Y., Wang, J., Dai, X., Wang, L., Yeh, C.-C. M., Zheng, Y., et al. (2023). How does attention work in vision transformers? A visual analytics attempt. *IEEE Transact. Vis. Comp. Graph*. doi: 10.1109/TVCG.2023.3261935

Lin, Y.-S., Chen, C.-C., and Greenlee, M. W. (2022). The role of lateral modulation in orientation-specific adaptation effect. *J. Vis*. 22, 13–13. doi: 10.1167/jov.22.2.13

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976–11986.

Lonnqvist, B., Bornet, A., Doerig, A., and Herzog, M. H. (2021). A comparative biology approach to dnn modeling of vision: a focus on differences, not similarities. *J. Vis*. 21, 17–17. doi: 10.1167/jov.21.10.17

Mahmood, K., Mahmood, R., and van Dijk, M. (2021). "On the robustness of vision transformers to adversarial examples," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7838–7847.

Martinez-Trujillo, J. (2022). Visual attention in the prefrontal cortex. *Ann. Rev. Vision Sci*. 8, 407–425. doi: 10.1146/annurev-vision-100720-031711

Moore, T., and Zirnsak, M. (2017). Neural mechanisms of selective visual attention. *Annu. Rev. Psychol*. 68, 47–72. doi: 10.1146/annurev-psych-122414-033400

Naseer, M., Ranasinghe, K., Khan, S., Hayat, M., Khan, F., and Yang, M.-H. (2021). Intriguing properties of vision transformers. *Adv. Neural Inf. Process. Syst*. 34, 23296–23308.

Nobre, A. C., Nobre, K., and Kastner, S. (2014). *The Oxford Handbook of Attention*. Oxford University Press.

Pan, Z., Zhuang, B., He, H., Liu, J., and Cai, J. (2022). "Less is more: Pay less attention in vision transformers," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36 (AAAI Press), 2035–2043.

Panaetov, A., Daou, K. E., Samenko, I., Tetin, E., and Ivanov, I. (2023). "Rdrn: recursively defined residual network for image super-resolution," in *Computer Vision-ACCV 2022*, eds L. Wang, J. Gall, T.-J. Chin, I. Sato, and R. Chellappa (Cham: Springer Nature Switzerland), 629–645.

Park, N., and Kim, S. (2022). "How do vision transformers work?," *In International Conference on Learning Representations*.

Pashler, H. (ed). (1998). *Attention, 1st Edn*. Psychology Press.

Paul, S., and Chen, P.-Y. (2022). Vision transformers are robust learners. *Proc. AAAI Conf. Artif. Intell*. 36, 2071–2081. doi: 10.1609/aaai.v36i2.20103

Peterson, M. A. (2015). "Low-level and high-level contributions to figure-ground organization," in *The Oxford Handbook of Perceptual Organization*, ed J. Wagemans (Oxford University Press), 259–280.

Poort, J., Raudies, F., Wannig, A., Lamme, V. A., Neumann, H., and Roelfsema, P. R. (2012). The role of attention in figure-ground segregation in areas v1 and v4 of the visual cortex. *Neuron* 75, 143–156. doi: 10.1016/j.neuron.2012.04.032

Qiu, F. T., Sugihara, T., and Von Der Heydt, R. (2007). Figure-ground mechanisms provide structure for selective attention. *Nat. Neurosci*. 10, 1492–1499. doi: 10.1038/nn1989

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks? *Adv. Neural Inf. Process. Syst*. 34, 12116–12128.

Ricci, M., Cadène, R., and Serre, T. (2021). Same-different conceptualization: a machine vision perspective. *Curr. Opin. Behav. Sci*. 37, 47–55. doi: 10.1016/j.cobeha.2020.08.008

RichardWebster, B., Anthony, S. E., and Scheirer, W. J. (2019). Psyphy: a psychophysics driven evaluation framework for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell*. 41, 2280–2286. doi: 10.1109/TPAMI.2018.2849989

Shushruth, S., Nurminen, L., Bijanzadeh, M., Ichida, J. M., Vanni, S., and Angelucci, A. (2013). Different orientation tuning of near-and far-surround suppression in macaque primary visual cortex mirrors their tuning in human perception. *J. Neurosci*. 33, 106–119. doi: 10.1523/JNEUROSCI.2518-12.2013

Srinivas, A., Lin, T.-Y., Parmar, N., Shlens, J., Abbeel, P., and Vaswani, A. (2021). "Bottleneck transformers for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16519–16529.

Stettler, D. D., Das, A., Bennett, J., and Gilbert, C. D. (2002). Lateral connectivity and contextual interactions in macaque primary visual cortex. *Neuron* 36, 739–750. doi: 10.1016/S0896-6273(02)01029-2

Styles, E. (2006). *The Psychology of Attention*. Psychology Press.

Sutherland, S. (1988). Feature selection. *Nature* 392, 350.

Tan, A., Nguyen, D. T., Dax, M., Nießner, M., and Brox, T. (2021). Explicitly modeled attention maps for image classification. *Proc. AAAI Conf. Artif. Intell*. 35, 9799–9807. doi: 10.1609/aaai.v35i11.17178

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H. (2021). "Training data-efficient image transformers and distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139, eds M. Meila, and T. Zhang (PMLR), 10347–10357.

Tsotsos, J. K. (1990). Analyzing vision at the complexity level. *Behav. Brain Sci*. 13, 423–445. doi: 10.1017/S0140525X00079577

Tsotsos, J. K. (2011). *A Computational Perspective on Visual Attention*. MIT Press.

Tsotsos, J. K. (2017). Complexity level analysis revisited: What can 30 years of hindsight tell us about how the brain might represent visual information? *Front. Psychol*. 8, 1216. doi: 10.3389/fpsyg.2017.01216

Tsotsos, J. K. (2022). When we study the ability to attend, what exactly are we trying to understand? *J. Imaging* 8, 212. doi: 10.3390/jimaging8080212

Tsotsos, J. K., Itti, L., and Rees, G. (2005). "A brief and selective history of attention," in *Neurobiology of Attention*, eds L. Itti, G. Rees, and J. K. Tsotsos (Academic Press).

Tsotsos, J. K., and Rothenstein, A. (2011). Computational models of visual attention. *Scholarpedia* 6, 6201. doi: 10.4249/scholarpedia.6201

Tuli, S., Dasgupta, I., Grant, E., and Griffiths, T. (2021). "Are convolutional neural networks or transformers more like human vision?," in *Proceedings of the Annual Meeting of the Cognitive Science Society, Vol. 43*.

Vaishnav, M., Cadene, R., Alamia, A., Linsley, D., VanRullen, R., and Serre, T. (2022). Understanding the computational demands underlying visual reasoning. *Neural Comput*. 34, 1075–1099. doi: 10.1162/neco_a_01485

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is All you need," in *Advances in Neural Information Processing Systems*, Vol. 30, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan,et al. (Curran Associates, Inc.)

Wloka, C., and Tsotsos, J. K. (2019). Flipped on its head: deep learning-based saliency finds asymmetry in the opposite direction expected for singleton search of flipped and canonical targets. *J. Vis*. 19, 318–318. doi: 10.1167/19.10.318

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). "Transformers: state-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (Association for Computational Linguistics), 38–45.

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., et al. (2021). "Visual transformers: where do transformers really belong in vision models?," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 599–609.

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., et al. (2021). "CvT: introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22–31.

Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollar, P., and Girshick, R. (2021). "Early convolutions help transformers see better,". in *Advances in Neural Information Processing Systems*, Vol. 34, eds M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan (Curran Associates, Inc.), 30392–30400.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., et al. (2015). "Show, attend and tell: neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37, F. Bach, and D. Blei (Lille: PMLR), 2048–2057.

Xu, Y., and Vaziri-Pashkam, M. (2021a). Examining the coding strength of object identity and nonidentity features in human occipito-temporal cortex and convolutional neural networks. *J. Neurosci*. 41, 4234–4252. doi: 10.1523/JNEUROSCI.1993-20.2021

Xu, Y., and Vaziri-Pashkam, M. (2021b). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nat. Commun*. 12, 2065. doi: 10.1038/s41467-021-22244-7

Xu, Y., and Vaziri-Pashkam, M. (2022). Understanding transformation tolerant visual object representations in the human brain and convolutional neural networks. *Neuroimage* 263, 119635. doi: 10.1016/j.neuroimage.2022.119635

Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., et al. (2021). "Focal attention for long-range interactions in vision transformers," in *Advances in Neural Information Processing Systems*, Vol. 34, eds M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan (Curran Associates, Inc.), 30008–30022.

Yantis, S., and Egeth, H. E. (1999). On the distinction between visual salience and stimulus-driven attentional capture. *J. Exp. Psychol*. 25, 661. doi: 10.1037/0096-1523.25.3.661

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.-H., et al. (2021). "Tokens-to-token ViT: training vision transformers from scratch on ImageNet," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 558–567.

Yue, X., Sun, S., Kuang, Z., Wei, M., Torr, P. H., Zhang, W., et al. (2021). "Vision transformer with progressive sampling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 387–396.

Zerroug, A., Vaishnav, M., Colin, J., Musslick, S., and Serre, T. (2022). "A benchmark for compositional visual reasoning," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhang, J., and Sclaroff, S. (2013). "Saliency detection: a boolean map approach," in *Proceedings of the IEEE International Conference on Computer Vision*, 153–160.

Zhou, D., Yu, Z., Xie, E., Xiao, C., Anandkumar, A., Feng, J., et al. (2022). "Understanding the robustness in vision transformers," in *Proceedings of the 39th International Conference on Machine Learning*, Vol. 162, eds K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato (PMLR), 27378–27394.

Zhou, H.-Y., Lu, C., Yang, S., and Yu, Y. (2021). "ConvNets vs. transformers: whose visual representations are more transferable?," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2230–2238.

Zhou, Q., Du, C., and He, H. (2022). Exploring the brain-like properties of deep neural networks: a neural encoding perspective. *Mach. Intell. Res.* 19, 439–455. doi: 10.1007/s11633-022-1348-x

Zhu, M., Hou, G., Chen, X., Xie, J., Lu, H., and Che, J. (2021). "Saliency-guided transformer network combined with local embedding for no-reference image quality assessment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1953–1962.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., et al. (2021). Unsupervised neural network models of the ventral visual stream. *Proc. Nat. Acad. Sci. U. S. A.* 118, e2014196118. doi: 10.1073/pnas.2014196118

Zucker, S. (1981). "Computer vision and human perception," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, 1102–1116.

Zucker, S. W. (1978). Vertical and horizontal processes in low level vision. *Comp. Vision Syst.* 187–195. Available online at: https://shop.elsevier.com/books/computer-vision-systems/hanson/978-0-12-323550-3

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 17 August 2023
DOI 10.3389/fpsyg.2023.1201681

OPEN ACCESS

EDITED BY
Mary A. Peterson,
University of Arizona, United States

REVIEWED BY
James R. Pomerantz,
Rice University, United States
James T. Townsend,
Indiana University Bloomington, United States
Rob Van Lier,
Radboud University, Netherlands

*CORRESPONDENCE
Ruth Kimchi
✉ rkimchi@univ.haifa.ac.il

RECEIVED 06 April 2023
ACCEPTED 31 July 2023
PUBLISHED 17 August 2023

CITATION
Kimchi R, Devyatko D and Sabary S (2023) Perceptual organization and visual awareness: the case of amodal completion.
Front. Psychol. 14:1201681.
doi: 10.3389/fpsyg.2023.1201681

Perceptual organization and visual awareness: the case of amodal completion

Ruth Kimchi[1,2]*, Dina Devyatko[2] and Shahar Sabary[2]

[1]Department of Psychology, University of Haifa, Haifa, Israel, [2]Institute of Information Processing and Decision Making, University of Haifa, Haifa, Israel

We investigated the involvement of visual awareness in amodal completion, and specifically, whether visual awareness plays a differential role in local versus global completion, using a primed shape discrimination paradigm and the color-opponent flicker technique to render the prime invisible. In four experiments, participants discriminated the shape of a target preceded by a partly occluded or a neutral prime. All primes were divergent occlusion patterns in which the local completion is based on good continuation of the contours at the point of occlusion and the global completion is based on maximum symmetry. The target corresponded to the shape that could arise as a result of local or global completion of the occluded prime. For each experiment with an invisible prime we conducted a version with a visible prime. Our results suggest that local completion, but not global completion, of a partly occluded shape can take place in the absence of visual awareness, but apparently only when the visible occluded shape generates a single, local completion. No completion, either local or global, appears to take place in the absence of visual awareness when the visible occluded shape generates multiple completions. The implications of these results to the differential role of visual awareness in local and global completions and to the relationship between multiple completions and unconscious amodal completions are discussed.

KEYWORDS

amodal completion, visual awareness, perceptual organization, global completion, local completion, symmetry, good continuation, color-opponent flicker (COF)

1. Introduction

Objects in our environment are often partly occluded by other objects or by themselves. Consequently, the input to our visual system is fragmented, yet we perceive our environment as a coherent scene with complete and whole objects. The visual system apparently fills in the incomplete parts of occluded objects, and does it rapidly and effortlessly. This filling in of contours and surfaces behind occluders has been referred to as amodal completion (Michotte et al., 1964,1991; see van Lier and Gerbino, 2015, for a review).

Two types of completions have been identified, local, and global. Local completion is based on local contour properties, mainly in line with the Gestalt principle of good continuation (Wertheimer, 1938) – a smooth connection between the visible contours of the occluded object (e.g., Kellman and Shipley, 1991; Wouterlood and Boselie, 1992; Fantoni and Gerbino, 2003). Global completion is based on global shape properties like symmetry and regularity (e. g., Buffart et al., 1981; Sekuler et al., 1994; van Lier et al., 1994, 1995b), fitting with the Gestalt Law of Prägnanz (Koffka, 1935).

In some occlusion patterns, referred to as convergent occlusion patterns, the local and global completions converge toward the same shape, whereas in the divergent occlusion patterns local and global completions yield different shapes (e.g., van Lier et al., 1995b; de Wit et al., 2006; Hazenberg et al., 2014). Research using various paradigms has demonstrated that indeed both local and global completions can be generated and the two completions can be competitive or one can prevail (Sekuler et al., 1994; van Lier et al., 1995a,b; van Lier and Wagemans, 1999; Plomp and van Leeuwen, 2006; Hazenberg et al., 2014).

Under certain conditions amodal completion appears to be cognitively impenetrable, as in the famous Kanizsa's "horse illusion" (Kanizsa, 1979), in which partly occluded horses tend to be interpreted as a single elongated horse even though it conflicts with our knowledge. Recent studies, however, demonstrated that amodal completion can be influenced by familiarity and knowledge (Hazenberg et al., 2014; Hazenberg and van Lier, 2016; Yun et al., 2018).

Here we aim to examine whether amodal completion can take place in the absence of visual awareness, and specifically, whether visual awareness plays a differential role in local versus global completion.

Research has recently addressed the question whether visual awareness of the stimulus is needed for it to be perceptually organized (e.g., Schwarzkopf and Rees, 2011; Montoro et al., 2014; Kimchi et al., 2018; Sabary et al., 2020). The results suggest that it depends on the perceptual organization processes under study, which is perhaps not surprising in light of the evidence that perceptual organization is a multiplicity of processes that vary in time course, developmental trajectory and attentional demands (e.g., Kimchi, 1998, 2000, 2009; Behrmann and Kimchi, 2003; Kimchi et al., 2005), and on the methods used to suppress the stimulus from awareness as they differ in the level at which the suppression takes place (e.g., Breitmeyer, 2015; Moors et al., 2016; Kimchi et al., 2018). Of particular relevance to the present article, grouping based on mirror symmetry was found to require visual awareness (Devyatko and Kimchi, 2020). Using a priming paradigm and a sandwich masking as an invisibility-inducing method, Devyatko and Kimchi presented participants with masked prime and a clearly visible target, which could be congruent or incongruent with the prime in symmetry. On each trial, the participants performed a two-alternative discrimination task on the target, and then rated the visibility of the prime on a subjective visibility four-point scale. Subjectively invisible primes failed to produce response priming, suggesting that symmetry detection may depend on visual awareness. This finding may suggest that global completion, to the extent that it is based on global symmetry, cannot take place in the absence of visual awareness. We note, however, that the stimuli in Devyatko and Kimchi's study were quite minimal – composed of just two vertical symmetric or asymmetric lines, thus having just one axis of symmetry when symmetrical. It is possible that in the presence of multiple axes of symmetry unconscious global completion can occur.

Also relevant to the present article are the findings from studies examining the relationship between visual awareness and another type of perceptual completion – modal completion. In modal completion the completed object has sensory qualities, as, for example, the Kanizsa's illusory triangle (Kanizsa, 1979), in which the observer perceives illusory contours and a bright surface in areas of the stimulus where there is no actual luminance discontinuity. There is little evidence that illusory contours can be formed in the absence of visual awareness. No perception of illusory contours was found when Kanizsa-type inducers were suppressed from awareness by binocular rivalry (Sobel and Blake, 2003), continuous flash suppression (CFS) (Harris et al., 2011), sandwich masking and counter-phase flickering (Banica and Schwarzkopf, 2016). In contrast, Wang et al. (2012), using breaking continuous flash suppression (b-CFS), found that a Kanizsa triangle emerged from suppression significantly faster than a control stimulus, presumably suggesting formation of illusory contours without awareness (but see Moors et al., 2016), and Jimenez et al. (2017) found priming by illusory figure masked by sandwich masking when the prime-mask SOA was 53 ms, but not when it was 23 ms. Thus, the results provide somewhat inconsistent evidence. Furthermore, the question whether amodal and modal completions share the same underlying mechanisms, as suggested by the "identity hypothesis" (Kellman and Shipley, 1991; Shipley and Kellman, 1992), or they have different mechanisms (Anderson et al., 2002; Singh, 2004; Anderson, 2007a), has been a matter of a furious debate (e.g., Kellman et al., 2007; Anderson, 2007b), and the controversy continues (see van Lier and Gerbino, 2015, for a discussion). Therefore we cannot draw clear predictions from the findings concerning the relationship between modal completion and visual awareness to the one between amodal completion and visual awareness.

To the best of our knowledge, Emmanouil and Ro's (2014) study is the only one to date that attempted to examine whether amodal completion can take place in the absence of visual awareness. Emmanouil and Ro examined the effect of invisible shape primes (a circle in Experiment 1 and a square in Experiment 2) on discrimination of a visible target. Invisibility was induced by metacontrast (Experiment 1) and backward masking (Experiment 2). They found that occluded and unoccluded primes produced a similar pattern of priming, suggesting that the invisible occluded primes were amodally completed. Note that the occluded patterns used by Emmanouil and Ro were highly familiar, convergent occlusion patterns, making it difficult to generalize their results to less familiar and to divergent occlusion patterns. More importantly, there are some concerns associated with Emmanouil and Ro's study, mainly regarding their prime identification task used for testing the visibility of the primes. The prime identification task in Experiments 1 and 2 included a target and the participants were instructed to ignore it. Not only could the to-be-ignored target bias perception of the prime, but identification of the prime could be susceptible to memory. Furthermore, in Experiment 1, a complete circle was considered as a "correct" response in the occluded and control prime conditions, thus actually testing whether the supposedly invisible prime was completed, rather than testing the visibility of the prime *per se*. In Experiment 2, the invisibility of the prime was questionable because prime identification was significantly above chance. Thus, the absence of a clear evidence for the invisibility of the prime in both experiments casts doubt on the interpretation of the observed results as indicating amodal completion without awareness.

In the present study, we used a priming paradigm in which participants discriminated the shape of a target preceded by a partly occluded or a neutral prime. The partly occluded primes were divergent occlusion patterns adapted from Sekuler et al. (1994) and Plomp and van Leeuwen (2006). The target corresponded to the shape that could arise as a result of a local or a global completion of the partly occluded prime. The prime was suppressed from awareness by

a modification of the color-opponent flicker (COF) method developed by Hoshiyama et al. (2006a,b). This method allows to present the prime for as long as required, and the luminance and contrast of the visual stimulus remain constant during the presentation period. This is important because amodal completion takes between 75 to 250 ms to complete, depending on the amount of occlusion and on the experimental task (Sekuler and Palmer, 1992; Murray et al., 2001; Guttman et al., 2003), so that backward masking commonly used to study unconscious processing, in which the prime is briefly presented (~40 ms), cannot be used. Thus, we presented the prime for 300 ms and ensured the invisibility of the prime during the presentation time.

Awareness of the prime was assessed by an objective visibility test, using a prime visibility task. Unconscious completion of the prime was measured as the difference between response to the target after the occluded prime and the neutral prime (*priming effect*). We reasoned that if the occluded prime is completed such that it is the same as the target, then performance after the occluded prime is expected to be better than after the neutral prime.

2. General methods

Each experiment included two parts. In the first part, participants performed the priming task; in the second part, following the completion of the first part, the participants performed the visibility task.

For each experiment with invisible prime we conducted a version with a visible prime to ensure that the primes and procedure that we used allow for amodal completion when the primes are visible.

2.1. Participants

Participants in all the experiments were students at the University of Haifa and were paid or granted a course credit for participation. All participants provided informed consent to a protocol approved by the Ethics Committee of the University of Haifa. All participants had normal vision and normal color vision and none, except for three, participated in more than one experiment. The sample size for the invisible experiments was calculated on the basis of an *a priori* power analysis (G*Power 3.1; Faul et al., 2007) to detect priming effects, given a moderate effect size (0.50), $\alpha = 0.05$ and 80% power. The sample size for the visible experiments was based on previously reported sample sizes in studies investigating amodal completion with priming paradigms (Sekuler et al., 1994; van Lier et al., 1995b; Hazenberg et al., 2014).

2.2. Apparatus

The experiments took place in a dimly lit room. All stimuli were generated using Matlab R2014a and Psychophysics Toolbox[1] and were presented on a 20″ sgi color monitor (C22BW711,1024 × 768 resolution, 100 Hz refresh rate) attached to a Mac Pro Late 2013 (3.7 GHz Quad-Core Intel Xeon ES). Responses were collected via Apple keyboard A1243emc 2,171. Participants viewed the stimuli at a distance of 57 cm with their head supported by a chin rest.

2.3. Stimuli

The prime stimuli were partly occluded shapes (occluded primes), and neutral primes comprised of two small squares (side: 0.25°) randomly placed within the area occupied by the partly occluded prime stimulus. The prime stimuli were drawn in red (R,G,B: 255,0,0; 13.8 cd/m2; x, y: 0.620, 0.348) and in green (R,G,B: 0,165,0; 13.8 cd/m2; x, y: 0.290, 0.594). When the color of the occluded shape was red, the color of the occluder was green, and when the color of the occluded shape was green the color of the occluder was red. Hereafter, the color of the occluded shape is used for referring to the color of the prime stimulus. The prime stimuli were drawn on a red-and-green checkerboard background (9° X 9°) and covered with a black mesh (Figure 1; see, Hoshiyama et al., 2006a,b). The average amount of contour and surface area occlusion was 20 to 25%.

All primes were divergent occlusion patterns; their local completion is always based on good continuation of the contours at the point of occlusion, and the global completion is based on maximum symmetry.

There were two types of targets corresponding to the two different shapes that could arise as a result of global and local completions of the occluded prime. The targets were presented on a grey (R,G,B: 170,170,170) background.

The stimuli for the visibility task corresponded to the global and local completions of the occluded primes with the occluder placed behind the figure.

The primes and targets in the priming task and the primes in the visibility task for Experiments 1–4 are presented in Figures 1A–D, respectively.

2.4. Procedure and design

2.4.1. Invisible prime experiments (Experiments 1a–4a)

These experiments were designed to examine whether amodal completion can take place in the absence of visual awareness. The prime was rendered invisible by means of a modification of the Color-Opponent Flicker (COF) method developed by Hoshiyama et al. (2006a,b). When two isoluminant opponent colors, for example, red and green, alternate at frequencies above the flicker fusion threshold (~30 Hz), the two colors fuse such that one uniformly yellow color is perceived (e.g., Schiller and Logothetis, 1990). In the conventional COF method, the red and green colors must be isoluminant. Hoshiyama et al.'s (2006a,b) modification of covering the figures with a black mesh prevents one color from being directly adjacent to another color such that no edges, caused by the difference in luminosity between two colors during COF, are produced, and consequently there is no need to strictly control the two colors for isoluminosity.

In each experiment, participants first undertook the masked priming task and following its completion they performed the visibility task.

1 http://psychtoolbox.org

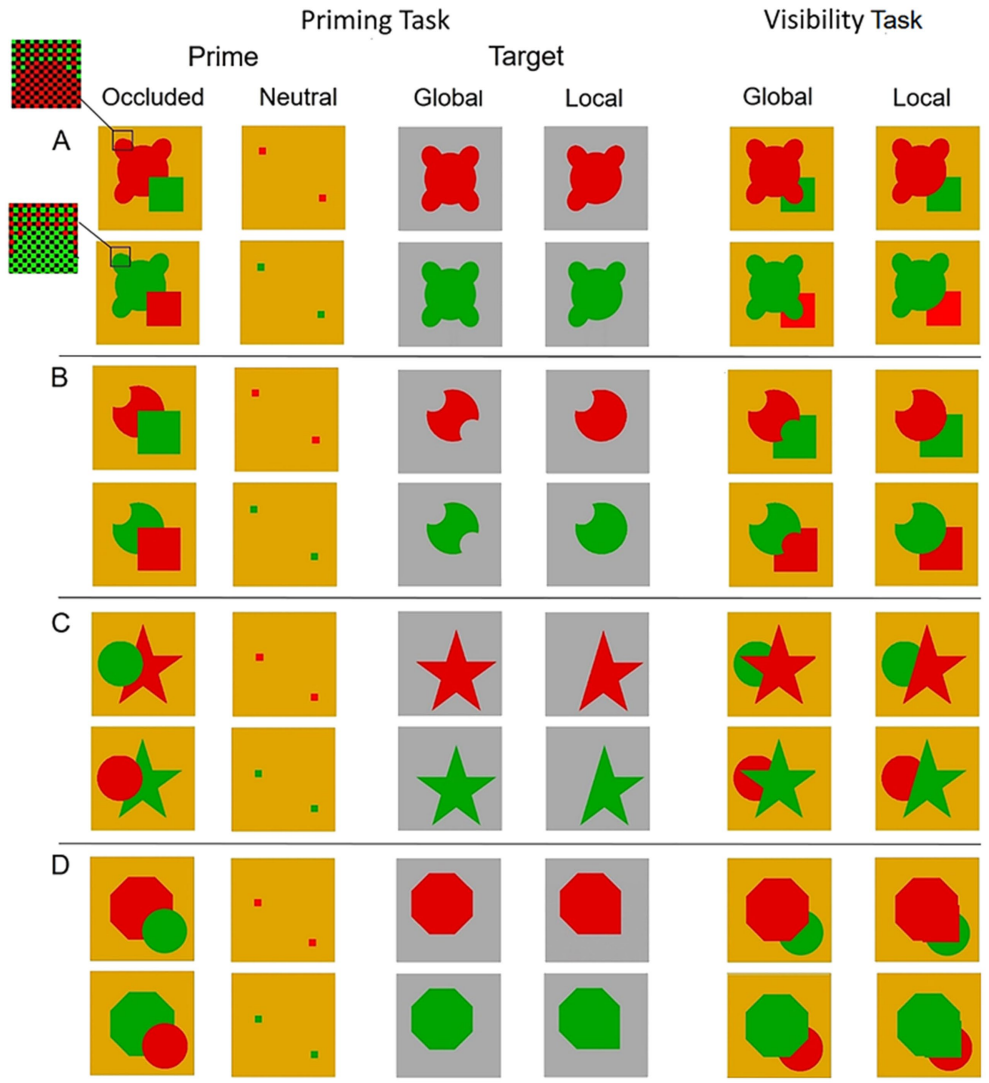


FIGURE 1
The primes and targets in the priming task, and the primes in the visibility task, used in **(A)** Experiment 1, **(B)** Experiment 2, **(C)** Experiment 3 and **(D)** Experiment 4. The primes were drawn in red or green on a red-and-green checkerboard background and covered with black mesh (see text for details).

2.4.1.1. Priming task

The sequence of events in a trial in the priming task is shown in Figure 2A (left panel). Each trial started with the presentation of a fixation mark (0.5° × 0.5° blue cross, R,G,B, 0,0,255) at the center of the screen for 1,000 ms. Then a pair of primes was presented alternately (e.g., red-green-red green…) at 100 Hz (10 ms presentation of each image) for 300 ms. Previous research have suggested that given the amount of occlusion that we used, presentation time of 300 ms is clearly sufficient for perceptual completion to take place (Murray et al., 2001; Guttman et al., 2003). The red and green primes fused and a uniform dark yellowish color was perceived. In half of the trials the alternated primes started with the red prime and in the other half with the green prime. Following the prime, a visible target (local or global) appeared and remained on the screen until the participant responded or 2,000 ms had elapsed. The color of the target shape was opposite to the color of the last prime during the COF sequence. Participants had to indicate the shape of the target by pressing one of two keys ("local shape" key or "global shape" key) with their dominant hand as fast as possible while avoiding making mistakes (Participants were told that the color of the shapes was irrelevant.)

All the combinations of prime type (occluded, neutral), first prime color in the alternated sequence (red, green) and target (global, local) were presented with equal frequency in a random order. Each participant completed 240 trials with five self-administrated breaks, preceded by 16 practice trials. During the practice, an auditory tone provided immediate feedback after an incorrect response or when 2,000 ms had elapsed with no response.

2.4.1.2. Visibility task

After completing the masked priming task participants performed the visibility task. The sequence of events in a trial, presented in Figure 2A (right panel), was similar to that of the priming task except that no target was presented after the presentation of the masked prime; instead of the target a question mark appeared

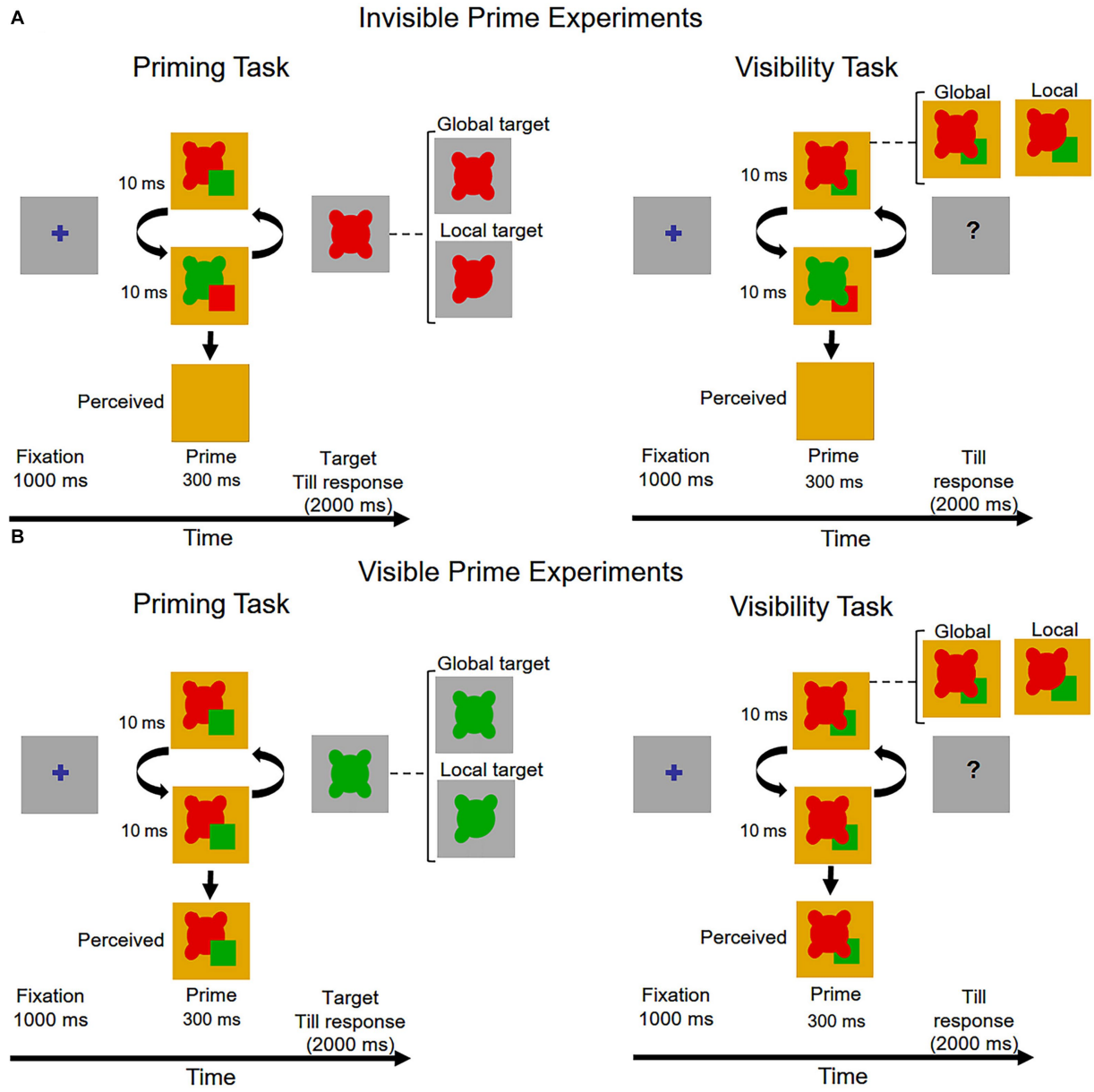


FIGURE 2
The sequence of events in a trial in the priming task (left panel) and in the visibility task (right panel) in **(A)** the invisible prime experiments and **(B)** the visible prime experiments. The figure depicts the prime and the global target of Experiment 1 (see text for details).

and stayed on the screen till response. The prime stimuli were a pair of red and green global primes or a pair of red and green local primes (see Figure 1). Participant had to indicate the shape of the prime by pressing one of two keys ("global shape" key or "local shape" key), and were instructed to guess the prime shape if the prime was invisible. All the combinations of the two prime types (global, local) and the first prime color (red, green) were presented with equal frequency in a random order. There were 40 trials preceded with 4 practice trials. No feedback was provided in the practice trials.

After completing the whole experiment, the participants were tested with two cards (# 70 and # 74) from Ishihara's Test for Color Deficiency in order to make sure they had normal color vision.

2.4.2. Visible prime experiments (Experiments 1b–4b)

For each of the invisible prime experiment we conducted a version in which the prime was visible. The visible prime experiment was similar to the corresponding invisible prime experiment, except that the colors of the primes did not alternate during a trial; either a red prime or a green prime was presented at 100 Hz (10 ms presentation of each image) for 300 ms, resulting in a visible prime. The sequence of events in a trial in the priming task and in the visibility task is presented in Figure 2B.

Apparatus, procedure and design were the same for all experiments.

2.5. Data analysis

All reaction time (RT) summaries and analyses are based on participants' mean RTs for correct responses. Trials with RTs shorter than 200 ms or longer than 1,600 ms were excluded from the analyses (less than 1% in all experiments). Because instructions to the participants emphasized both speed and accuracy, and to simplify presentation and analyses, we used an inverse efficiency (IE) score (mean correct RT divided by proportion of correct responses) as the dependent measure (Townsend and Ashby, 1978, 1983). Using IE in the present data analyses was appropriate given the high accuracy rate, which exceeded 94% in all our experiments, and the similar pattern of results for RT and accuracy measures (Bruyer and Brysbaert, 2011; Vandierendonck, 2017, 2018). Repeated measures ANOVAs were used to analyze the IE data. All ANOVAs were calculated using SAS (version 9.4). See the Supplementary materials for analyses of accuracy and RT separately.

When null effects were theoretically important, i.e., inferring that awareness may be necessary for amodal completion to occur, we also evaluated evidence in favor of the null hypothesis by computing the Bayes factor (BF10) in a Bayesian paired t-test, using JASP statistical software (www.jasp-stats.org) and a Cauchy prior centered on zero (scale = 0.707).

3. Experiment 1

3.1. Participants

Twenty-seven individuals (24 females and 3 males, 4 left-handed, age ranged 18–29, $M = 23.2$) participated in Experiment 1a (Invisible-prime experiment), and 18 individuals (13 females and 5 males, 3 left-handed, age ranged 19–31, $M = 23.1$) participated in Experiment 1b (Visible-prime experiment).

3.2. Stimuli

The basic shape was a symmetrical shape consisting of a large circle (3.95° in diameter) and four elliptical protrusions and containing two axes of symmetry. The overall size of the shape was 4.8° X 4.2°. The symmetrical shape was partly occluded by a 2.7° × 2.7° square, constituting the occluded prime (Figure 1A). The global completion of the occluded shape resulted in the symmetrical shape, whereas the local completion by good continuation of the visible contours of the occluded shape resulted in a different shape with three elliptical protrusions and one axis of symmetry. The two targets corresponded to the two shapes that resulted from global and local completions, respectively (Figure 1A). The centers of both targets were moved 0.53° below and to the right from the center of the occluded prime in order to avoid full overlapping of prime's and targets' contours. Two primes were used in the visibility task, designated as global and local, which were produced by placing the occluding square behind the shapes generated by the global and local completions (Figure 1A).

3.3. Results and discussion

3.3.1. Experiment 1a: invisible prime

Trials in which RT was longer than 1,600 ms or shorter than 200 ms were excluded from the analyses (0.34%). Mean RTs and mean accuracy (AC) for global and local targets in the neutral and occluded prime conditions are presented in Table 1. Mean IE scores are presented in Figure 3A.

Performance in the visibility task did not differ significantly from chance, mean accuracy = 51.85%, $t(26) = 1.43$, $p = 0.1665$, indicating that COF rendered the prime invisible.

The repeated measures ANOVA with prime (neutral and occluded) and target (global and local) as within-subject factors, conducted on the IE scores, revealed a main effect of target, $F(1,26) = 10.11$, $p = 0.0038$, $\eta_p^2 = 0.28$. As can be seen in Figure 3A, performance for the global target was better than performance for the local target, both in the neutral prime condition, $t(26) = 2.972$, $p = 0.003$, Cohen's d = 0.572, and in the occluded prime condition, $t(26) = 2.421$, $p = 0.011$, Cohen's d = 0.455.

The main effect of prime, $F(1,26) = 1.78$, $p = 0.1941$, and the interaction between target and prime conditions, $F < 1$, were not significant, showing no indication of priming effects. Bayesian paired t-tests showed that the evidence provides substantial support for the null hypothesis for global priming, BF10 = 0.275, and is inconclusive for local priming, BF10 = 0.742.

These results are seen to suggest that no completion, either local or global, took place when the occluded prime was invisible.

TABLE 1 Mean RTs (ms) and mean AC (%) for global and local target as a function of prime condition (neutral and occluded) for each experiment.

	Global target				Local target			
	Neutral prime		Occluded prime		Neutral prime		Occluded prime	
	RT	AC	RT	AC	RT	AC	RT	AC
Experiment 1a	540.89	97.46	539.18	97.59	555.88	94.75	553.41	96.17
Experiment 1b	581.82	97.96	561.67	96.94	584.18	97.15	567.17	97.31
Experiment 2a	536.62	95.61	548.49	97.03	550.26	95.92	542.28	96.54
Experiment 2b	545.63	96.85	546.54	95.37	542.75	96.48	521.46	95.83
Experiment 3a	471.56	97.71	472.26	97.28	466.74	97.59	472.09	97.35
Experiment 3b	492.15	96.57	483.71	96.94	489.76	96.85	472.49	97.50
Experiment 4a	471.27	98.02	474.39	98.63	468.91	97.84	469.29	97.22
Experiment 4b	504.96	97.68	492.64	98.24	498.63	97.02	502.37	96.66

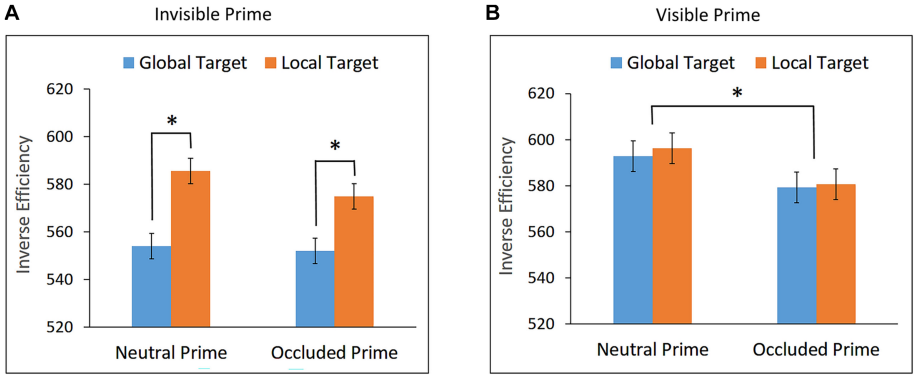


FIGURE 3
Inverse Efficiency (IE) scores for global and local targets in the neutral and occluded prime conditions in **(A)** Experiment 1a – Invisible prime, and **(B)** Experiment 1b – Visible prime. Error bars represent within subjects ± SEM.

3.3.2. Experiment 1b: visible prime

Trials in which RT was shorter than 200 ms or longer than 1,600 ms were excluded from the analyses (0.65%). Mean RTs and mean AC for global and local targets in the neutral and occluded prime conditions are presented in Table 1. Mean IE scores are presented in Figure 3B.

Performance in the visibility task was significantly above chance, mean accuracy = 96.94%, $t(26) = 71.49$, $p < 0.0001$, confirming that the prime was visible.

The repeated measures ANOVA conducted on the IE data showed no main effect of target, $F < 1$. The effect of prime was significant, $F(1,17) = 4.84$, $p = 0.0419$, $\eta_p^2 = 0.22$, indicating that performance was better when the targets appeared after the occluded prime than after the neutral prime. As can be seen in Figure 3B, this facilitation effect did not interact with target, $F < 1$, suggesting global and local priming effects, which did not differ significantly in magnitude.

These results suggest that when the partly occluded prime was visible, both global and local completions were generated and no completion was significantly preferred over the other. Multiple completion were previously observed with different stimuli (van Lier et al., 1995a,b), and van Lier et al. suggested that the preference for a global or a local completion is the consequence of a competition between interpretations.

Interestingly, Sekuler et al. (1994) used a stimulus similar to ours and a primed matching paradigm, and found a clear global completion. What may account for the discrepancy in the results? First, the stimuli used by Sekuler and colleagues were "square like," having 4 axes of symmetry, whereas the stimuli we used were elongated and had only 2 axes of symmetry. Second, although the stimuli were very similar, they could differ in the amount of occlusion. In the absence of specific details we can just eyeball and it seems that the amount of contour occlusion in Sekuler et al. was somewhat larger, such that the connection between the visible contours of the occluded object was smoother in our stimulus, increasing the likelihood of local completion; thus a local completion was generated and competed with the global completion. Third, the presentation of the prime was different, and so was the task (target shape discrimination vs. same-different judgments). These differences could in principle affect the pattern of results, but further research is required in order to understand how.

In contrast to the multiple completions observed for the visible prime, the results for the invisible prime (Experiment 1a) showed no completion, either local or global, suggesting that amodal completion cannot take place in the absence of visual awareness.

4. Experiment 2

4.1. Participants

Twenty-seven individuals (17 females and 10 males, 3 left-handed, age ranged 18–34, $M = 24.1$) participated in Experiment 2a, and 18 individuals (11 females and 7 males, 3 left-handed, age ranged 19–30, $M = 22.9$) participated in Experiment 2b.

4.2. Stimuli

The basic shape was a symmetrical shape generated by a large circle (4.5° in diameter) with a circular cut-off in the upper left side and in the upper right side (2.25° in diameter each), the center of which located at the large circle circumference. The shape contained two axes of symmetry. The symmetrical shape was partly occluded by a 3.75° × 3.75° square, constituting the occluded prime (Figure 1B). The global completion of the occluded shape resulted in the symmetrical shape. The local completion by good continuation of the visible contours of the occluded shape resulted in a different shape of a circle with a single cut-off in the upper left side, and contained one axis of symmetry. The two targets corresponded to the two shapes that resulted from global and local completions (Figure 1B). The centers of the large circles in both targets were 0.8° below the center of the prime's large circle in order to avoid full overlapping of prime's and targets' contours. Two primes were used in the visibility task, global and local, which were produced by placing the occluding square behind the global and local completed shapes (Figure 1B).

4.3. Results and discussion

4.3.1. Experiment 2a: invisible prime

Trials with RTs shorter than 200 ms or longer than 1,600 ms were excluded from the analysis (0.45%). Mean RTs and mean accuracy for

global and local targets in the neutral and occluded prime conditions are presented in Table 1. Mean IE scores are presented in Figure 4A.

Performance in the visibility task was at chance, mean accuracy = 51.76%, $t = 0.90$, $p = 0.3777$, confirming that the prime was invisible.

The repeated measures ANOVA conducted on the IE data showed no main effects of target, $F < 1$, and prime, $F < 1$. The interaction between target and prime, however, was significant, $F(1,26) = 5.15$, $p = 0.0318$, $\eta_p^2 = 0.17$. As can be seen in Figure 4A, performance for the local target was better following the occluded prime than the neutral prime, $t(26) = 1.949$, $p = 0.031$, Cohen's $d = 0.375$, indicating local priming. No such effect whatsoever was observed for the global target, $t(26) = -0.889$, $p = 0.809$; a Bayesian paired t-test showed that the evidence provides substantial support for the null hypothesis, BF10 = 0.117.

To rule out the possibility that the observed local priming was due to the performance of participants for whom the prime was visible, we calculated, for each participant, the local priming score and the visibility score (i.e., accuracy in the visibility task) and examined whether there is a positive correlation between these two scores. The analysis yielded no significant correlation, $r = -0.284$, $p = 0.924$; Bayesian correlation showed that the evidence provided substantial support for the null hypothesis, BF10 = 0.106.

These results suggest that local completion of a partly occluded object can take place in the absence of visual awareness.

4.3.2. Experiment 2b: visible prime

Trials with RTs shorter than 200 ms or longer than 1,600 ms were excluded from the analysis (0.83%). Mean RTs and mean accuracy for global and local targets in the neutral and occluded prime conditions are presented in Table 1. Mean IE scores are presented in Figure 4B.

Performance in the visibility task was significantly above chance, mean accuracy = 94.03%, $t(17) = 30.76$, $p < 0.0001$, confirming that the prime was visible.

The repeated measures ANOVA conducted on the IE data showed no significant main effects of prime, $F < 1$, nor of target, $F(1,17) = 1.57$, $p = 0.2274$. The interaction between target and prime, however, was significant, $F(1,17) = 7.78$, $p = 0.0127$, $\eta_p^2 = 0.31$. As can be seen in Figures 4B, a priming effect for the local target was observed, $t(17) = 3.028$, $p = 0.004$, Cohen's $d = 0.717$, suggesting local completion. No priming effect was observed for the global target, $t(17) = -0.715$, $p = 0.758$; Bayesian paired t-test showed that the evidence provides substantial support for the null hypothesis, BF10 = 0.155.

In contrast to our results, Plomp and van Leeuwen (2006) found some preference for global completion for a similar stimulus. However, the procedures of their study and the current study are quite different, making the comparison between the two studies difficult.

Inspection of our Figure 4 shows that the pattern of results with the invisible prime (Experiment 2a) was similar to the one with the visible prime (Experiment 2b): a facilitation for the response to the local target following occluded prime, and no such facilitation whatsoever for the response to the global target. These results suggest that local completion was taking place both in the presence and in the absence of visual awareness.

Comparing the results of Experiments 1 and 2 reveals an interesting pattern. In Experiment 1, no completion was observed for the invisible prime (Experiment 1a), which, when visible, generated multiple completions (Experiment 1b). In Experiment 2, local completion was observed for the invisible prime (Experiment 2a), which, when visible generated a single local completion (Experiment 2b). Presumably, the potential generation of multiple completions versus a single completion may influence whether or not unconscious completion occurs. We return to this issue later.

5. Experiment 3

Previous research showed, as noted earlier, that familiarity and knowledge can have an effect on amodal completion (Hazenberg et al., 2014; Hazenberg and van Lier, 2016; Yun et al., 2018). Experiment 3 was designed to examine whether familiarity can influence amodal completion in the absence of awareness. To this end, the prime we used was a partially occluded five-point star (Figure 1C). Familiarity in this case favors the global completion.

5.1. Participants

Twenty-seven individuals (18 females and 9 males, 1 left-handed, age ranged 19–34, M = 25.85) participated in Experiment 3a, and 18

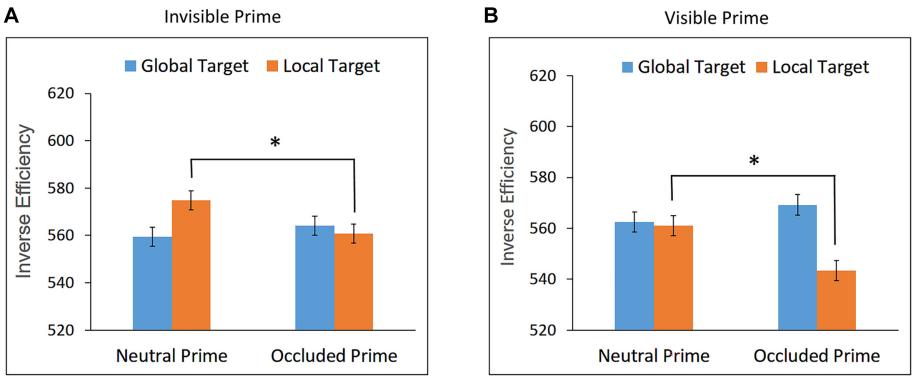


FIGURE 4
Inverse Efficiency scores for global and local targets as a function of prime condition (occluded and neutral) in **(A)** Experiment 2a – Invisible prime, and **(B)** Experiment 2b – Visible prime. Error bars represent within subjects ± SEM.

individuals (13 females and 5 males, 2 left-handed, age ranged 19–35, $M = 23.89$) participated in Experiment 3b.

5.2. Stimuli

The basic shape was a five-point star (6.2° by 6.2° in height and width), containing five axes of symmetry. The star was partly occluded by a circle (3.5° in diameter), constituting the occluded prime (Figure 1C). The global completion of the occluded shape resulted in the star. The local completion by good continuation of the collinear visible contours of the occluded star resulted in a different shape, and contained one axis of symmetry. The two targets corresponded to the two shapes that resulted from the global and local completions (Figure 1C). The centers of both targets were moved 0.5° below and to the right from the center of the occluded prime in order to avoid full overlapping of prime's and targets' contours. Two primes were used in the visibility task, global and local, which were produced by placing the occluding circle behind the global and local completed shapes (Figure 1C).

5.3. Results and discussion

5.3.1. Experiment 3a: invisible prime

Trials with RTs shorter than 200 ms or longer than 1,600 ms were excluded from the analysis (0.22%). Mean RT and mean accuracy for global and local targets in the neutral and occluded prime conditions are presented in Table 1. Mean IE scores are presented in Figure 5A.

Performance in the visibility task did not differ from chance, mean accuracy = 50.37%, $t(26) = 0.34$, $p = 0.7343$, confirming that the prime was invisible.

None of the effects tested by the repeated measures ANOVA, conducted on the IE scores, reached statistical significance: $F < 1$, $F(1,26) = 3.53$, $p = 0.0714$, and $F < 1$, for the effects of target, prime and prime X target interaction, respectively. These results show no indication of global or local priming; Bayesian paired t-tests showed that the evidence provides substantial support for the null hypothesis for the former, BF10 = 0.133, and strong support for the latter, BF10 = 0.080.

These findings indicate that neither local nor global completion took place when the occluded prime was invisible, and suggest that familiarity had no influence on amodal completion in the absence of visual awareness.

5.3.2. Experiment 3b: visible prime

Trials with RTs shorter than 200 ms or longer than 1,600 ms were excluded from the analysis (0.53%). Mean RTs and mean accuracy for global and local targets in the neutral and occluded prime conditions are presented in Table 1. Mean IE scores are presented in Figure 5B.

Performance in the visibility task was significantly above chance, mean accuracy = 96.53%, $t(17) = 50.07$, $p < 0.0001$, confirming that the prime was visible.

The repeated measures ANOVA showed no significant effect of target, $F < 1$. The effect of prime was significant, $F(1,17) = 11.58$, $p = 0.0034$, $\eta_p^2 = 0.41$, indicating better performance for the targets following an occluded prime than the neutral prime. As can be seen in Figure 5B, this facilitation effect did not interact with target, $F < 1$, suggesting both global and local priming effects, which did not differ significantly in magnitude.

Interestingly, although both familiarity and maximum symmetry favor global completion of the partly occluded star, no preference for global completion was observed in our experiment. The local completion constituted a competing alternative, presumably due to the collinearity of the lines at the point of occlusion. We note that Plomp and van Leeuwen (2006) found some preference for global completion for the star stimulus. However, as noted earlier, the procedures of their study and the current study are quite different, making the comparison between the two studies difficult.

The results of Experiment 3 are similar to the ones of Experiment 1, indicating no completion, either local or global, in the absence of visual awareness for an occluded shape, which when visible generated multiple completions. On the other hand, local completion in the absence of visual awareness was observed for an occluded shape that when visible generated a single local completion (Experiment 2).

No indication of unconscious global completion was found. However, the results of Experiment 2 suggest that in order to reach a clear conclusion regarding the necessity of visual awareness for

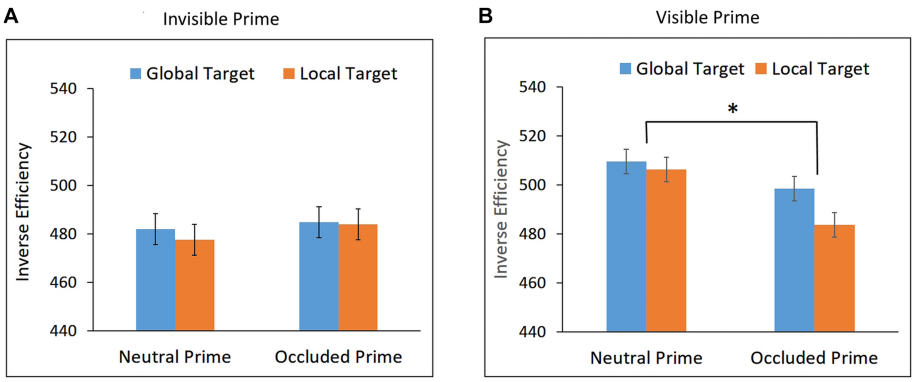


FIGURE 5
Inverse Efficiency scores for global and local targets as a function of prime condition (occluded and neutral) in **(A)** Experiment 3a – Invisible prime, and **(B)** Experiment 3b – Visible prime. Error bars represent within subjects ± SEM.

global completion to occur, unconscious global completion has to be examined with an occluded shape that generates a single global completion when visible. The next experiment was designed to do so.

6. Experiment 4

This experiment was designed to examine whether global completion can take place in the absence of visual awareness. To this end we used an octagon partly occluded by a circle. We suspected that the occluded octagon could be a good candidate for generating a single global completion because the amount of symmetry axes – eight axes in the octagon versus only one in the locally completed shape – strengthen the tendency for global completion, whereas the relatively weak connection between the lines at the point of occlusion (the lines meet at 90° angle) weakens the tendency for local completion. We ran first the visible prime version of this experiment, the results of which confirmed our supposition, and then we ran the invisible version. To be consistent with all other experiments reported in this article, we keep the same order, reporting first the invisible prime experiment and then the visible prime one.

6.1. Participants

Twenty-seven individuals (19 females and 8 males, 3 left-handed, age ranged 19–36, M = 25.67) participated in Experiment 4a, and 18 individuals (15 females and 3 males, 1 left-handed, age ranged 18–41, $M = 24.72$) participated in Experiment 4b.

6.2. Stimuli

The basic shape was an octagon (4.7° × 4.7° in height and width), containing eight axes of symmetry. The octagon was partly occluded by a circle (3.5° in diameter), constituting the occluded prime (Figure 1D). The global completion of the occluded shape resulted in the octagon. The local completion by continuation of the visible contours of the occluded octagon resulted in a different shape, and contained one axis of symmetry. The two targets corresponded to the two shapes that resulted from the global and local completions (Figure 1D). The centers of both targets were moved 0.5° above and to the left from the center of the occluded prime in order to avoid full overlapping of prime's and targets' contours. Two primes were used in the visibility task, global and local, which were produced by placing the occluding circle behind the global and local completed shapes (Figure 1D).

6.3. Results and discussion

6.3.1. Experiment 4a: invisible prime

Trials with RTs shorter than 200 ms or longer than 1,600 ms were excluded from the analysis (0.15%). Mean RT and mean accuracy for global and local targets in the neutral and occluded prime conditions are presented in Table 1. Mean IE scores are presented in Figure 6A.

Performance in the visibility task did not differ from chance, mean accuracy = 49.72%, $t(26) = -0.21$, $p = 0.832$, confirming that the prime was invisible.

The repeated measures ANOVA showed no significant effect of target, prime, and prime X target interaction, Fs < 1, indicating no local or global priming. Bayesian paired t-tests showed that the evidence provides substantial support for the null hypothesis, BF10 = 0.116, BF10 = 0.207, for local and global priming, respectively. These findings suggest that neither local nor global completion took place when the occluded prime was invisible.

6.3.2. Experiment 4b: visible prime

Trials with RTs shorter than 200 ms or longer than 1,600 ms were excluded from the analysis (0.39%). Mean RTs and mean accuracy for global and local targets in the neutral and occluded prime conditions are presented in Table 1. Mean IE scores are presented in Figure 6B.

Performance in the visibility task was significantly above chance, mean accuracy = 97.36%, $t(17) = 61.6$, $p < 0.0001$, confirming that the prime was visible.

The repeated measures ANOVA showed no significant effect of target, $F < 1$, nor of prime, $F < 1$. The interaction between target and prime, however, was significant, $F(1,17) = 5.04$, $p = 0.0383$, $\eta_p^2 = 0.23$. As can be seen in Figures 6B, a priming effect for the global target was observed, $t(17) = 2.099$, $p = 0.026$, Cohen's d = 0.495, suggesting global completion. No priming effect was observed for the local target, $t(17) = -0.594$, $p = 0.720$; Bayesian paired t-test showed that the evidence provides substantial support for the null hypothesis, BF10 = 0.165.

These results indicate that when the partly occluded octagon was visible, a single global completion was generated. In contrast, no perceptual completion was observed when the partly occluded octagon was invisible.

Thus, the results of Experiment 4 suggest that no global completion can take place in the absence of visual awareness. Assuming that the global completion is based on the overall symmetry of the occluded shape, this finding is in agreement with the finding reported by Devyatko and Kimchi's (2020) that symmetry-based grouping requires visual awareness, although the stimuli in their study had one vertical axes of symmetry whereas the present octagon had eight axes of symmetry.

In addition, the results of Experiment 4 suggest that generating a single completion is not sufficient for unconscious completion to occur.

7. General discussion

In this study we investigated whether amodal completion can take place in the absence of visual awareness, and specifically, whether visual awareness plays a differential role in local versus global completion. To this end we used a primed shape discrimination paradigm in which the prime was rendered invisible by color-opponent flicker (COF; Hoshiyama et al., 2006a,b). All primes were divergent occlusion patterns in which the local completion is always based on good continuation of the contours at the point of occlusion, and the global completion is based on maximum symmetry. The targets corresponded to the two different shapes that could arise as a result of global and local

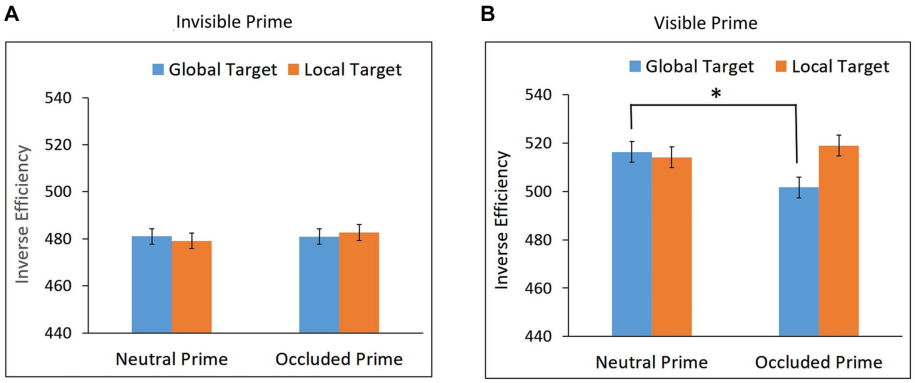


FIGURE 6
Inverse Efficiency scores for global and local targets as a function of prime condition (occluded and neutral) in **(A)** Experiment 4a – Invisible prime, and **(B)** Experiment 4b – Visible prime. Error bars represent within subjects ± SEM.

completions of the occluded prime. For each of the invisible prime experiment we conducted a version in which the prime was visible.

The results provide a somewhat complicated picture. No significant local or global priming was observed for invisible occluded primes that generated multiple completions when visible (Experiments 1 and 3), suggesting that no completion, either global or local, can take place in the absence of visual awareness for occluded shapes that, when visible, generate both local and global completions. A significant local priming for an invisible occluded prime was found, suggesting that local completion can occur in the absence of visual awareness, but only for an occluded prime that when visible generates a single local completion (Experiment 2). In contrast, the results showed no global priming for an invisible occluded prime, suggesting that no global completion can take place in the absence of visual awareness, even when the occluded prime generates a single global completion when visible (Experiment 4). In addition, familiarity did not have an effect on unconscious amodal completion (Experiment 3). We note however, that in our study familiarity favored global completion. It would be interesting to examine the influence of familiarity when it favors local completion.

Taken together, the results of Experiments 1–4 are seen to have two important implications. One concerns the role of visual awareness in local versus global completion. The other concerns the relationship between potential multiple completions and unconscious amodal completion. The two are not completely unrelated.

The present results demonstrate that visual awareness plays a differential role in local versus global completion: local completion can take place in the absence of visual awareness, whereas visual awareness is required for global completion to occur. This is perhaps not surprising. Local completion is based mainly on the basic grouping principle of good continuation (Wertheimer, 1938), which plays an important role in contour integration (e.g., Field et al., 1993; Kimchi, 2000; Geisler et al., 2001), processed in the early brain regions V1 and V2 (e.g., Roelfsema, 2006), and appears to operate in the absence of visual awareness (Breitmeyer et al., 2005; Devyatko et al., 2019). Global completion, is based on maximum symmetry. In contrast to good continuation, symmetry produces strong responses only in higher-order regions, especially V4 and LOC (e.g., Sasaki et al., 2005), symmetry-based grouping was found to require visual awareness (Devyatko and Kimchi, 2020; but see, Makin et al., 2023), and the role of symmetry in perceptual organization is not entirely clear because of its interaction with other grouping factors (van der Helm, 2015). For example, a number of researchers argue that grouping by other factors precedes and facilitates grouping by symmetry (see for discussion Machilsen et al., 2009). Also, it was found that organization by collinearity alone suffices for automatic capture of attention by a perceptual object, whereas organization by symmetry alone does not, suggesting that symmetry may play a weaker role than collinearity in the formation of objecthood (Kimchi et al., 2016). Furthermore, according to the Gestaltists, symmetry (like closure) is not a grouping factor *per se*, but rather it plays a critical role in how the perceptual system arrives at a stable, organized structure, with symmetry being particularly crucial in determining figural goodness (Koffka, 1935; Wertheimer, 1938; Palmer, 1991). It is possible that this can be achieved only when the stimulus is consciously perceived.

One may argue that the differential role of visual awareness in local versus global completion supports the view that local and global completions are qualitatively different processes (e. g., Kellman et al., 2005). This view suggests that local completion is a bottom-up process based on stimulus structure whereas global completion is a top-down process, referred to by Carrigan et al. (2016) and Kellman et al. (2005) as "recognition from partial information" (but see Peta et al., 2019 for a critical discussion).

In our opinion, however, the present results do not necessarily suggest that there is a qualitative difference between the processes involved in local and global completions, as both processes can be based on stimulus properties, which can be simple local properties or more complex global properties. This of course does not rule out the possibility of top-down influences such as familiarity and knowledge on amodal completion (e.g., Hazenberg et al., 2014; Yun et al., 2018). But familiarity and knowledge are not to be confused with symmetry and regularity, because the formers depend on the perceiver's past experience whereas the latter on stimulus structure (see also Peta et al., 2019). Our results are seen to suggest that completion based on simple, local properties can take place in the absence of visual awareness, at least under certain conditions, but completion based on more complex global properties such as symmetry cannot.

The second implication of the present results concerns the relationship between potential multiple completions and unconscious amodal completion. It appears that when there is an unresolved

competition between local and global completions, no completion can take place in the absence of visual awareness. Our results also show that a single completion is not sufficient for unconscious completion to occur. Obviously, further research is required in order to get a clearer picture of this relationship. The results of our visible prime experiments show, on the one hand, either a local or a global completion (Experiments 2 and 4), and on the other hand, two competing completions without a local or a global preference (Experiments 1 and 3). It would be interesting to find out what happens in the absence of visual awareness when there is competition between the two completions, but the competition is resolved and one completion prevails. Namely, the question is whether it is the mere presence of a competition or the presence of unresolved competition that requires visual awareness for amodal completion to take place.

Before concluding, we note that unconscious global and local completions need to be explored with different suppression methods, because previous research showed that the extent of information processing without consciousness is also dependent on the invisibility-inducing method – i.e., on the level at which the suppression induced by the method takes place (e.g., Breitmeyer, 2015; Moors et al., 2016; Kimchi et al., 2018).

To summarize, our results suggest that local completion, but not global completion, of a partly occluded shape can take place in the absence of visual awareness, but apparently only when the visible occluded shape generates a single, local completion. No completion appears to take place in the absence of visual awareness when the visible occluded shape generates multiple completions. Further research is required to clarify the relationship between multiple completions and unconscious amodal completion, as well as the effect of familiarity on unconscious completion.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the Ethics Committee of the University of Haifa. The participants provided their written informed consent to participate in this study.

Author contributions

RK: conceptualization, methodology, formal analysis, writing, supervision, funding acquisition. DD and SS: methodology, testing and data collection, formal analysis. All authors approved the submitted version.

Funding

This research was supported by a grant (grant number 1473/15) from the Israel Science Foundation (ISF) to RK.

Acknowledgments

We thank Dan Manor for his help in programming the experiments, Mor Leder, Alisa Kanterman and Noura Khoury for their help in running the experiments, and the reviewers for their helpful comments.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1201681/full#supplementary-material

References

Anderson, B. L. (2007a, 2005). The demise of the identity hypothesis and the insufficiency and nonnecessity of contour relatability in predicting object interpolation: comment on Kellman, Garrigan, and Shipley. *Psychol. Rev.* 114, 470–487. doi: 10.1037/0033-295X.114.2.470

Anderson, B. L. (2007b). Filling-in models of completion: rejoinder to Kellman, Garrigan, Shipley, and Keane (2007) and Albert (2007). *Psychol. Rev.* 114, 509–525. doi: 10.1037/0033-295X.114.2.509

Anderson, B. L., Singh, M., and Fleming, R. (2002). The interpolation ofobject and surface structure. *Cogn. Psychol.* 44, 148–190. doi: 10.1006/cogp.2001.0765

Banica, T., and Schwarzkopf, D. S. (2016). Induction of Kanizsa Contours Requires Awareness of the Inducing Context. *PLoS One* 11:e0161177. doi: 10.1371/journal.pone.0161177

Behrmann, M., and Kimchi, R. (2003). What does visual agnosia tell us about perceptual organization and its relationship to object perception? *J. Exp. Psychol. Hum. Percept. Perform.* 29, 19–42. doi: 10.1037/0096-1523.29.1.19

Breitmeyer, B. G. (2015). Psychophysical 'blinding' methods reveal a functional hierarchy of unconscious visual processing. *Conscious. Cogn. Int. J.* 35, 234–250. doi: 10.1016/j.concog.2015.01.012

Breitmeyer, B. G., Ogmen, H., Ramon, J., and Chen, J. (2005). Unconscious and conscious priming by forms and their parts. *Vis. Cogn.* 12, 720–736. doi: 10.1080/13506280444000472

Bruyer, R., and Brysbaert, M. (2011). Combining speed and accuracy in cognitive psychology: Is the inverse efficiency score (IES) a better dependent variable than the mean reaction time (RT) and the percentage of errors (PE)? *Psychol. Bel.* 51:5. doi: 10.5334/pb-51-1-5

Buffart, H., Leeuwenberg, E., and Restle, F. (1981). Coding theory of visual pattern completion. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 241–274. doi: 10.1037/0096-1523.7.2.241

Carrigan, S. B., Palmer, E. M., and Kellman, P. J. (2016). Differentiating global and local contour completion using a dot localization paradigm. *J. Exp. Psychol. Hum. Percept. Perform.* 42, 1928–1946. doi: 10.1037/xhp0000233

de Wit, T. C., Bauer, M., Oostenveld, R., Fries, P., and van Lier, R. (2006). Cortical responses to contextual influences in amodal completion. *NeuroImage* 32, 1815–1825. doi: 10.1016/j.neuroimage.2006.05.008

Devyatko, D., and Kimchi, R. (2020). Visual awareness is essential for grouping based on mirror symmetry. *Symmetry* 12:1872. doi: 10.3390/sym12111872

Devyatko, D., Sabary, S., and Kimchi, R. (2019). Perceptual organization of line configurations: Is visual awareness necessary? *Conscious. Cogn. Int. J.* 70, 101–115. doi: 10.1016/j.concog.2019.02.005

Emmanouil, T. A., and Ro, T. (2014). Amodal completion of unconsciously presented objects. *Psychon. Bull. Rev.* 21, 1188–1194. doi: 10.3758/s13423-014-0590-9

Fantoni, C., and Gerbino, W. (2003). Contour interpolation by vector-field combination. *J. Vis.* 3, 281–303. doi: 10.1167/3.4.4

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). GPower 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/bf03193146

Field, D. J., Hayes, A., and Heiss, R. F. (1993). Contour integration by the human visual system: Evidence for a local "association field.". *Vis. Res.* 33, 173–193. doi: 10.1016/0042-6989(93)90156-Q

Geisler, W. S., Perry, J. S., Super, B. J., and Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vis. Res.* 41, 711–724. doi: 10.1016/S0042-6989(00)00277-7

Guttman, S. E., Sekuler, A. B., and Kellman, P. J. (2003). Temporal Variations in Visual Completion: A Reflection of Spatial Limits? *J. Exp. Psychol. Hum. Percept. Perform.* 29, 1211–1227. doi: 10.1037/0096-1523.29.6.1211

Harris, J. J., Schwarzkopf, D. S., Song, C., Bahrami, B., and Rees, G. (2011). Contextual illusions reveal the limit of unconscious visual processing. *Psychol. Sci.* 22, 399–405. doi: 10.1177/0956797611399293

Hazenberg, S. J., Jongsma, M., Koning, A., and van Lier, R. (2014). Differential familiarity effects in amodal completion: Support from behavioral and electrophysiological measurements. *J. Exp. Psychol. Hum. Percept. Perform.* 40, 669–684. doi: 10.1037/a0034689

Hazenberg, S. J., and van Lier, R. (2016). Disentangling effects of structure and knowledge in perceiving partly occluded shapes: An ERP study. *Vis. Res.* 126, 109–119. doi: 10.1016/j.visres.2015.10.004

Hoshiyama, M., Kakigi, R., Takeshima, Y., Miki, K., and Watanabe, S. (2006a). Priority of face perception during subliminal stimulation using a new color-opponent flicker stimulation. *Neurosci. Lett.* 402, 57–61. doi: 10.1016/j.neulet.2006.03.054

Hoshiyama, M., Kakigi, R., Takeshima, Y., Miki, K., and Watanabe, S. (2006b). Differential priming effects of color-opponent subliminal stimulation on visual magnetic responses. *Hum. Brain Mapp.* 27, 811–818. doi: 10.1002/hbm.20222

Jimenez, M., Montoro, P. R., and Luna, D. (2017). Global shape integration and illusory form perception in the absence of awareness. *Conscious. Cogn.* 53, 31–46. doi: 10.1016/j.concog.2017.05.004

Kanizsa, G. (1979). *Organization in vision: Essays on Gestalt psychology*. New York, NY: Praeger Publishers.

Kellman, P. J., Garrigan, P., and Shipley, T. F. (2005). Object Interpolation in Three Dimensions. *Psychol. Rev.* 112, 586–609. doi: 10.1037/0033-295X.112.3.586

Kellman, P. J., Garrigan, P., Shipley, T. F., and Keane, B. P. (2007). Interpolation processes in object perception: Reply to Anderson (2007). *Psychol. Rev.* 114, 488–502. doi: 10.1037/0033-295x.114.2.488

Kellman, P. J., and Shipley, T. F. (1991). A theory of visual interpolation in object perception. *Cogn. Psychol.* 23, 141–221. Retrieved from. doi: 10.1016/0010-0285(91)90009-D

Kimchi, R. (1998). Uniform connectedness and grouping in the perceptual organization of hierarchical patterns. *J. Exp. Psychol. Hum. Percept. Perform.* 24, 1105–1118. doi: 10.1037/0096-1523.24.4.1105

Kimchi, R. (2000). The perceptual organization of visual objects: A microgenetic analysis. *Vis. Res.* 40, 1333–1347. doi: 10.1016/s0042-6989(00)00027-4

Kimchi, R. (2009). Perceptual organization and visual attention. *Prog. Brain Res.* 176, 15–33. doi: 10.1016/S0079-6123(09)17602-1

Kimchi, R., Devyatko, D., and Sabary, S. (2018). Can perceptual grouping unfold in the absence of awareness? Comparing grouping during continuous flash suppression and sandwich masking. *Conscious. Cogn.* 60, 37–51. doi: 10.1016/j.concog.2018.02.009

Kimchi, R., Hadad, B., Behrmann, M., and Palmer, S. E. (2005). Microgenesis and Ontogenesis of Perceptual Organization: Evidence From Global and Local Processing of Hierarchical Patterns. *Psychol. Sci.* 16, 282–290. doi: 10.1111/j.0956-7976.2005.01529.x

Kimchi, R., Yeshurun, Y., Spehar, B., and Pirkner, Y. (2016). Perceptual organization, visual attention, and objecthood. *Vis. Res.* 126, 34–51. doi: 10.1016/j.visres.2015.07.008

Koffka, K. (1935). *Principles of Gestalt Psychology*. New York: Harcourt Brace Jovanovich.

Machilsen, B., Pauwels, M., and Wagemans, J. (2009). The role of vertical mirror symmetry in visual shape detection. *J. Vis.* 9, 11.1–11.1111. doi: 10.1167/9.12.11

Makin, A. D. J., Roccato, M., Karakashevska, E., Tyson-Carr, J., and Bertamini, M. (2023). Symmetry Perception and Psychedelic Experience. *Symm. Percept. Psych. Exp. Symm.* 15:1340. doi: 10.3390/sym15071340

Michotte, A., Thines, G., Costall, A., and Butterworth, G. (1991). *Michotte's experimental phenomenology of perception*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Michotte, A., Thines, G., and Crabbe, G. (1964). *Les compliments amodaux des structures perceptives*. Oxford, England: Publications U. Louvain.

Montoro, P. R., Luna, D., and Ortells, J. J. (2014). Subliminal Gestalt grouping: evidence of perceptual grouping by proximity and similarity in absence of conscious perception. *Conscious. Cogn.* 25, 1–8. doi: 10.1016/j.concog.2014.01.004

Moors, P., Wagemans, J., van Ee, R., and de-Wit, L. (2016). No evidence for surface organization in Kanizsa configurations during continuous flash suppression. *Atten. Percept. Psychophys.* 78, 902–914. doi: 10.3758/s13414-015-1043-x

Murray, R. F., Sekuler, A. B., and Bennett, P. J. (2001). Time course of amodal completion revealed by a shape discrimination task. *Psychon. Bull. Rev.* 8, 713–720. doi: 10.3758/BF03196208

Palmer, S. E. (1991). "Goodness, Gestalt, groups, and Garner: Local symmetry subgroups as a theory of figural goodness" in *The perception of structure: Essays in honor of Wendell R. Garner*. eds. G. R. Lockhead and J. R. Pomerantz (Washington, DC, US: American Psychological Association)

Peta, A., Fantoni, C., and Gerbino, W. (2019). Mid-level priming by completion vs. mosaic solutions. *I-Perception* 10:204166951882034. doi: 10.1177/2041669518820347

Plomp, G., and van Leeuwen, C. (2006). Asymmetric priming effects in visual processing of occlusion patterns. *Percept. Psychophys.* 68, 946–958. doi: 10.3758/bf03193357

Roelfsema, P. R. (2006). Cortical Algorithms for Perceptual Grouping. *Annu. Rev. Neurosci.* 29, 203–227. doi: 10.1146/annurev.neuro.29.051605.112939

Sabary, S., Devyatko, D., and Kimchi, R. (2020). The role of visual awareness in processing of global structure: Evidence from the perceptual organization of hierarchical patterns. *Cognition* 205:104442. doi: 10.1016/j.cognition.2020.104442

Sasaki, Y., Vanduffel, W., Knutsen, T., Tyler, C., and Tootell, R. (2005). Symmetry activates extrastriate visual cortex in human and nonhuman primates. *Proc. Natl. Acad. Sci. U. S. A.* 102, 3159–3163. doi: 10.1073/pnas.0500319102

Schiller, P. H., and Logothetis, N. K. (1990). The color-opponent and broad-band channels of the primate visual system. *Trends Neurosci.* 13, 392–398. doi: 10.1016/0166-2236(90)90117-s

Schwarzkopf, D. S., and Rees, G. (2011). Interpreting local visual features as a global shape requires awareness. *Proc. R. Soc. B Biol. Sci.* 278, 2207–2215.

Sekuler, A. B., and Palmer, S. E. (1992). Perception of partly occluded objects: A microgenetic analysis. *J. Exp. Psychol. Gen.* 121, 95–111. doi: 10.1037/0096-3445.121.1.95

Sekuler, A. B., Palmer, S. E., and Flynn, C. (1994). Local and global processes in visual completion. *Psychol. Sci.* 5, 260–267. doi: 10.1111/j.1467-9280.1994.tb00623.x

Shipley, T. F., and Kellman, P. J. (1992). Perception of partly occluded objects and illusory figures: Evidence for an identity hypothesis. *J. Exp. Psychol. Hum. Percept. Perform.* 18, 106–120.

Singh, M. (2004). Modal and amodal completion generate different shapes. *Psychol. Sci.* 15, 454–459. doi: 10.1111/j.0956-7976.2004.00701.x

Sobel, K. V., and Blake, R. (2003). Subjective contours and binocular rivalry suppression. *Vis. Res.* 43, 1533–1540. doi: 10.1016/S0042-6989(03)00178-0

Townsend, J. T., and Ashby, F. G. (1978). "Methods of modeling capacity in simple processing systems" in *Cognitive theory*. eds. J. Castellan and F. Restle (Hillsdale, N.J: Erlbaum)

Townsend, J. T., and Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes*. Cambridge Cambridgeshire; New York: Cambridge University Press.

van der Helm, P. A. (2015). "Symmetry perception" in *The Oxford Handbook of Perceptual Organization*. ed. J. Wagemans (Oxford: Oxford University Press)

van Lier, R., and Gerbino, W. (2015). "Perceptual completions" in *Oxford handbook of perceptual organization*. ed. J. Wagemans (Oxford: Oxford University Press)

van Lier, R., Leeuwenberg, E., and van der Helm, P. (1995a). Multiple completions primed by occlusion patterns. *Perception* 24, 727–740. doi: 10.1068/p240727

van Lier, R., van der Helm, P., and Leeuwenberg, E. (1994). Integrating global and local aspects of visual occlusion. *Perception* 23, 883–903. doi: 10.1068/p230883

van Lier, R., van der Helm, P., and Leeuwenberg, E. (1995b). Competing global and local completions in visual occlusion. *J. Exp. Psychol. Hum. Percept. Perform.* 21, 571–583.

van Lier, R., and Wagemans, J. (1999). From images to objects: Global and local completions of self-occluded parts. *J. Exp. Psychol. Hum. Percept. Perform.* 25, 1721–1741.

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behav. Res. Methods* 49, 653–673. doi: 10.3758/s13428-016-0721-5

Vandierendonck, A. (2018). Further tests of the utility of integrated speed-accuracy measures in task switching. *J. Cogn.* 1:8. doi: 10.5334/joc.6

Wang, L., Weng, X., and He, S. (2012). Perceptual grouping without awareness: superiority of Kanizsa triangle in breaking interocular suppression. *PLoS One* 7:e40106. doi: 10.1371/journal.pone.0040106

Wertheimer, M. (1938). "Laws of organization in perceptual forms" in *A source book of Gestalt psychology*. ed. W. D. Ellis (London: Routledge & Kegan Paul) (Original work published 1923)

Wouterlood, D., and Boselie, F. (1992). A good-continuation model of some occlusion phenomena. *Psychol. Res.* 54, 267–277. doi: 10.1007/bf01358264

Yun, X., Hazenberg, S. J., and van Lier, R. (2018). Temporal properties of amodal completion: Influences of knowledge. *Vis. Res.* 145, 21–30. doi: 10.1016/j.visres.2018.02.011

frontiers | Frontiers in Computer Science

TYPE Technology and Code
PUBLISHED 13 September 2023
DOI 10.3389/fcomp.2023.1140723

OPEN ACCESS

EDITED BY
Ernest Greene,
University of Southern California, United States

REVIEWED BY
William McIlhagga,
University of Bradford, United Kingdom
Eckart Michaelsen,
System Technologies and Image Exploitation IOSB, Germany

*CORRESPONDENCE
Dirk B. Walther
✉ dirk.bernhardt.walther@utoronto.ca

†These authors have contributed equally to this work

RECEIVED 09 January 2023
ACCEPTED 18 August 2023
PUBLISHED 13 September 2023

CITATION
Walther DB, Farzanfar D, Han S and Rezanejad M (2023) The mid-level vision toolbox for computing structural properties of real-world images.
Front. Comput. Sci. 5:1140723.
doi: 10.3389/fcomp.2023.1140723

The mid-level vision toolbox for computing structural properties of real-world images

Dirk B. Walther*, Delaram Farzanfar†, Seohee Han† and Morteza Rezanejad

Department of Psychology, University of Toronto, Toronto, ON, Canada

Mid-level vision is the intermediate visual processing stage for generating representations of shapes and partial geometries of objects. Our mechanistic understanding of these operations is limited, in part, by a lack of computational tools for analyzing image properties at these levels of representation. We introduce the Mid-Level Vision (MLV) Toolbox, an open-source software that automatically processes low- and mid-level contour features and perceptual grouping cues from real-world images. The MLV toolbox takes vectorized line drawings of scenes as input and extracts structural contour properties. We also include tools for contour detection and tracing for the automatic generation of vectorized line drawings from photographs. Various statistical properties of the contours are computed: the distributions of orientations, contour curvature, and contour lengths, as well as counts and types of contour junctions. The toolbox includes an efficient algorithm for computing the medial axis transform of contour drawings and photographs. Based on the medial axis transform, we compute several scores for local mirror symmetry, local parallelism, and local contour separation. All properties are summarized in histograms that can serve as input into statistical models to relate image properties to human behavioral measures, such as esthetic pleasure, memorability, affective processing, and scene categorization. In addition to measuring contour properties, we include functions for manipulating drawings by separating contours according to their statistical properties, randomly shifting contours, or rotating drawings behind a circular aperture. Finally, the MLV Toolbox offers visualization functions for contour orientations, lengths, curvature, junctions, and medial axis properties on computer-generated and artist-generated line drawings. We include artist-generated vectorized drawings of the Toronto Scenes image set, the International Affective Picture System, and the Snodgrass and Vanderwart object images, as well as automatically traced vectorized drawings of set architectural scenes and the Open Affective Standardized Image Set (OASIS).

KEYWORDS

mid-level vision, perceptual grouping, gestalt grouping rules, contour drawings, medial axis transform, symmetry, contour tracing

Introduction

Visual processing relies on different transformations of a visual representation derived from the pattern of light on the retina. In the early stages of visual processing, primary visual cortex (V1) encodes a representation of natural scene statistics based on contrast, orientation, and spatial frequencies (Hubel and Wiesel, 1962; Olshausen and Field, 1996; Vinje and Gallant,

2000). In later stages of visual processing, the semantic content of a visual scene is encoded in scene-selective regions based on category information (Epstein and Kanwisher, 1998; Epstein et al., 2001). Yet, despite a mechanistic understanding of these representations, we know less about the intervening stages of visual processing.

Mid-level vision is the intermediate visual processing stage for combining elementary features into conjunctive feature sets representing shapes and partial geometries of objects and scenes (Peirce, 2015; Malcolm et al., 2016). Along the ventral visual pathway, anatomical regions V2, V3, and V4 are the likely biological substrate supporting these operations, whose contributions to visual processing are far less understood (Peirce, 2015). Some evidence suggests that V2 is sensitive to border ownership, and V4 encodes curvature and symmetry information (Peterhans and von der Heydt, 1989; Gallant et al., 1996; Pasupathy and Connor, 2002; Wilder et al., 2022). Drawing from physiologically plausible representations of mid-level visual processing, we offer a set of computational tools for image processing to help fill this gap in our understanding of visual perception and help uncover intermediate stages of visual processing. Understanding mid-level vision allows us to investigate how discrete percepts are constructed and used to facilitate goal-driven behaviors.

Much of mid-level vision operations are qualitatively explained by Gestalt psychology (Koffka, 1935). Gestalt grouping cues are principles of perceptual organization that explain how basic visual elements are organized into meaningful whole percepts – these principles are proximity, similarity, continuity, closure, and figure/ground (Wertheimer, 1922). Empirical studies of Gestalt grouping cues frequently use stylized lab stimuli, such as clouds of dots (e.g., Kubovy, 1994; Wagemans, 1997; Norcia et al., 2002; Sasaki, 2007; Bona et al., 2014, 2015), Gabor patches (e.g., Field et al., 1993; Machilsen et al., 2009), or simple shapes (e.g., Elder and Zucker, 1993; Wagemans, 1993; De Winter and Wagemans, 2006). Typically, these stimuli are *constructed* to contain a specific amount of symmetry, contour integration, parallelism, closure, etc. By comparison, little empirical work has been done on testing Gestalt grouping principles for perceiving complex, real-world scenes (but see Geisler et al., 2001; Elder and Goldberg, 2002). More recent research in human and computer vision has extended the work of Wertheimer to physiologically plausible representations of shapes using the medial axis transform (Blum, 1967; Ayzenberg and Lourenco, 2019; Rezanejad et al., 2019, 2023; Ayzenberg et al., 2022).

Underlying medial axis-based representations of shape is an understanding of vision in terms of contours and shapes. Contours and shapes form the basis of early theories of vision, such as Marr's 2 ½ D sketch (Marr and Nishihara, 1978; Marr, 1982), or the recognition- by-components model (Biederman, 1987), as well as practical applications to the recognition of three-dimensional objects (Lowe, 1987). Perceptual organization is recognized to play an important role in these systems (Feldman and Singh, 2005; Lowe, 2012; Pizlo et al., 2014) as well as in computer vision more generally (Desolneux et al., 2004, 2007; Michaelsen and Meidow, 2019).

We here provide a software toolbox[1] for the study of mid-level vision using naturalistic images. This toolbox opens an avenue for testing mid-level features' role in visual perception by *measuring* low- and mid-level image properties from contour drawings and real-world photographs. Our measurement techniques are rooted in biologically inspired computations for detecting geometric relationships between contours. Working on contour geometry has the clear advantage of resulting in tractable, mechanistic algorithms for understanding mid-level vision. However, it has the disadvantage of not being computable directly from image pixels. We overcome this difficulty by offering algorithms that detect contours in color photographs and trace the contours to arrive at vectorized representations.

Contour extraction

Most functions in the MLV Toolbox rely on vectorized contour drawings. These drawings can be generated by humans tracing the important contours in photographs, or by importing existing vector graphics from an SVG file with **importSVG**, or by automatically detecting edges from the photographs and tracing the contours in the extracted edge maps.

Edge detection is performed using a structured forest method known as the Dollár edge detector (Dollár and Zitnick, 2013, 2014). We here use the publicly available Structured Edge Detection Toolbox V3.0.[2] This computationally efficient edge detector achieves excellent accuracy by predicting local edge masks in a structured learning framework applied to random decision forests. As the code for this toolbox was written in Matlab, this software became a natural choice as the edge detector for our toolbox. Using image-specific adaptive thresholding, we generate a binarized version from the edge map and its corresponding edge strength. The binarized edge map is then morphologically thinned to create one-pixel-wide contour segments.

Our method for tracing contours is adapted from the Edge Linking and Line Segment Fitting code sections from Peter Kovesi's Matlab and Octave Functions for Computer Vision and Image Processing.[3] These are edge-linking functions that enable the system to take a binary edge image and create lists of connected edge points. Additional helper functions fill in small gaps in a given binary edge map image (edgelink) and form straight line segments to sets of line segments that are shorter than a given tolerance value (lineseg).

Functions: traceLineDrawingFromRGB, **traceLinedrawingFromEdgeMap**

The definition of the beginning and end of contours depends on the method of generation. We provide two data sets, for which the contours were drawn by trained artists using a graphics tablet. For these data sets, the beginning of a contour is defined as the artist putting the pen on the graphics tablet and the end by them lifting the pen up. For automatically traced contours, the beginning and end are defined by the beginning and end of lists of connected edge points.

Both methods result in vectorized line drawings that are represented as a set of contours (Figure 1). Each contour consists of a list of successive, connected straight line segments. The information is contained in a vecLD struct with the following fields:

1 http://mlvtoolbox.org

2 https://github.com/pdollar/edges

3 https://www.peterkovesi.com/matlabfns/#edgelink

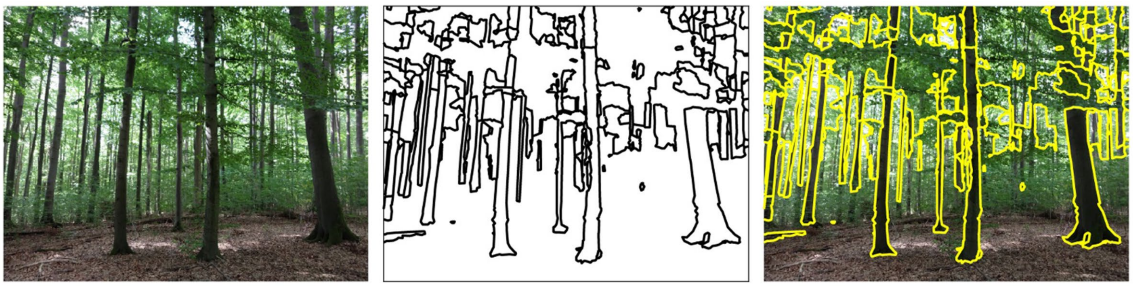


FIGURE 1
Color photograph of a forest (left), extracted contours (middle), contours superimposed on the original image (right).

originalImage:	the file name of the original photograph.
imsize:	[width, height].
lineMethod:	a descriptive string indicating how the line drawing was generated, e.g., 'artist', 'importSVG', 'traceLineDrawingFromRGB'.
numContours:	the number of contours.
contours:	cell array of size (1, numContours) containing the individual contour information. Each cell is an N x 4 array. Each row of the array represents one contour line segment. The columns are the start and end coordinates of the line segments in the order: X1, Y1, X2, Y2. Note that the end point of one segment is the start point of the next segment. This way of storing the coordinates is somewhat redundant, but it allows for greater efficiency for plotting, processing, and splitting contours.

More fields are added to the struct as image properties are computed. Vectorized line drawings can be plotted into a figure window using the **drawLinedrawing** function. They can be rendered into a binary image using the **renderLinedrawing** function.

Measuring contour properties

Analysis of properties of individual contours and contour segments follows the intuitive definitions outlined below.

Orientation of individual contours is computed as:

$$ori = \arctan\left(\frac{Y2 - Y1}{X2 - X1}\right)$$

where orientations are measured in degrees in the counterclockwise direction, starting from 0° at horizontal all the way to 360° back at horizontal. Orientations are stored in a direction-specific way so that 180° is not considered the same as 0°. This coding is important for computing curvature and junction angles correctly.

When computing histograms of orientation, however, orientation angles are computed modulo 180 degrees. Orientation histograms are weighted by the number of pixels (length) of a particular line segment. By default, eight histograms are computed with bin centers at 0, 22.5, 45, 67.5, 90, 112.5, 135, and 157.5 degrees.

Functions: **computeOrientation**, **getOrientationStats**

The **length** of contour segments is computed as the Euclidean distance between the start and end points:

$$length = \sqrt{(X2 - X1)^2 + (Y2 - Y1)^2}$$

The length of an entire contour is the sum of the lengths of the individual contour segments. Contour histograms are computed with bins equally spaced on a logarithmic scale within the bounds of 2 pixels and (width + height). For instance, an eight-bin histogram (the default) for images of size 800 × 600 pixels has bin centers located at 3.4, 8.5, 19.5, 43.2, 94.2, 204.2, 441.5, and 953.1 pixels.

Functions: **computeLength**, **getLengthStats**

Mathematically, **curvature** is defined as the change in angle per unit length. In calculus, the change of angle is given by the second derivative. For the piecewise straight line segments in our implementation, we compute the curvature for each line segment as the amount of change in orientation from this to the next line segment, divided by the length of the segment:

$$curvature_i = \frac{\mathrm{mod}(ori_{i+1} - ori_i, 180)}{length_i}$$

For the last segment of a contour, we use the angle difference with the previous instead of the next segment. Histograms of contour curvature are computed with bins equally spaced on a logarithmic scale between 0 degrees / pixel (straight line; no curvature) and 90 degrees / pixel (a minimal line of 2 pixels length making a sharp 180-degree turn). For a default eight-bin histogram, bins are centered at 0.33, 1.33, 3.09, 6.20, 11.65, 21.23, 38.06, and 67.64 degrees per pixel.

Functions: **computeCurvature**, **getCurvatureStats**

Contour junctions are detected at the intersections between contours. The intersections are computed algebraically from the coordinates of all contour segments. Intersections of contour segments within a contour are not considered. Junctions are still detected when contours do not meet exactly. This function is controlled by two parameters, a relative epsilon (RE) and an absolute epsilon (AE). The relative epsilon controls the allowable gap between the end point of a contour segment and the hypothetical junction location as the fraction

of the length of the contour segment. The absolute epsilon determines the maximum allowable gap in pixels. The minimum between the two gap measures is used as a threshold value for detecting junctions. Hand tuning of the parameters resulted in values of AE = 1 pixel and RE = 0.3. These two parameters can be set as optional arguments of the **detectJunctions** function.

Contour segments participating in junctions are algorithmically severed into two new segments at the junction locations whenever junctions are detected far enough away from the end points. The angles between all segments participating in a junction are measured as the difference in (directed) orientation angle between adjacent segments. Inspired by previous literature on contour junctions (Biederman, 1987), junction types are classified according to how many contour segments participate in the junction and according to their angles as follows:

3 segments:	determine the maximum angle ± between any two segments.
	Y junctions: $\alpha < 160°$.
	T junctions: $160° \leq \alpha \leq 200°$, i.e., $\alpha = 180° \pm 20°$.
	Arrow junctions:$\alpha > 200°$.
4 segments:	*X junctions*
>4 segments:	*Star junctions*

Junctions between two contour segments are sometimes described as L junctions. Here, we do not consider L junctions as they would occur at every location where one contour segment ends and the next begins, making them too numerous to be useful.

Integer counts of the number of junctions of each count are collected in junction histograms, which can optionally be normalized by the total number of pixels in a vectorized line drawing. The minimum angles between any of the contour segments, which are bounded between 0 and 120 degrees, are also counted as "Junction Angles" in a histogram with bin centers at 7.5, 22.5, 37.5, 52.5, 67.5, 82.5, 97.5, and 112.5 degrees.

Functions: **computeJunctions**, **getJunctionStats**, **detectJunctions**

For each contour property, the following fields are added to the vecLD struct:

property:	(1, numContours) cell array; each cell contains a vector of properties for the corresponding contour segments
propertyHistograms:	(numContours, numBins) array with the property histograms for the individual contours
normPropertyHistograms:	the same but normalized by the total number of contour pixels
sumPropertyHistogram:	(1,numBins) array: the property histogram for the entire image – the sum of propertyHistograms
normSumPropertyHistogram:	the same but normalized by the total number of contour pixels
propertyBins:	(1,numBins) array: the centers of the histogram bins

To visualize the contour properties, use the **drawLinedrawingProperty** function (Figure 2). The first argument to the function is a vecLD struct, the second a string denoting the contour property. The function draws the color drawing into the current figure window. Use **drawAllProperties** to visualize all contour properties for a given vecLD struct, either using subplots or in separate figure windows. The **renderLinedrawingProperty** function draws the contour properties into an image instead of a figure.

We have used these contour properties previously to explain behavior (Walther and Shen, 2014; Wilder et al., 2018) and neural mechanisms of scene categorization (Choo and Walther, 2016). We have also related statistics of these properties to the emotional content of scenes and generated artificial images with specific property combinations to elicit emotional responses (Damiano et al., 2021a), as well as to esthetic pleasure (Farzanfar and Walther, 2023).

Medial axis-based properties

Blum (1967) was probably the first to introduce medial axis-based representations and the method for producing them using a grassfire analogy. Imagine a shape cut out of a piece of paper set on fire around its border, where the fire front moves toward the center of the shape at a constant pace. Skeletal points are formed at the locations where the fire fronts collide. In other words, we can look at the Medial Axis Transform (MAT) as a method for applying the grassfire process to disclose its quench sites and associated radius values (Feldman and Singh, 2006). The MAT provides a complete visual representation of shapes, as it is applicable to all bounded shapes as well as the areas outside of closed shapes. Humans have been shown to rely on the shape skeleton defined by the MAT when they attend to objects (Firestone and Scholl, 2014), represent shapes in memory (Ayzenberg and Lourenco, 2019), or judge the esthetic appeal of shapes (Sun and Firestone, 2022).

In this toolbox, we compute measures of the relationships between contours using the medial axis transform. Visually, the medial axis transform is made up of a number of branches that come together at branch points to create a shape skeleton. A group of contiguous regular points from the skeleton that are located between two junction points, two end points, or an end point and a junction point are known as skeletal branches. The behavior of the average outward flux (AOF) of the gradient of the Euclidean distance function to the boundary of a 2D object through a shrinking disk can be used to identify skeletal points that lie on skeletal branches and identify the types of those three classes of points, as demonstrated by Rezanejad (2020). We go over this calculation in the following.

Imagine that the distance transform D_Ω of a shape Ω is a signed distance function that indicates the closest distance of a given point $\mathbf{p}$ to the shape's boundary $\partial\Omega$ (Figure 3A). Formally, we can imagine a positive sign for the distance value when $\mathbf{p}$ is inside the shape Ω and a negative sign when $\mathbf{p}$ is outside of the shape Ω. We can then define the distance function gradient vector for point $\mathbf{p}$ as $\dot{\mathbf{q}}_\Omega(\mathbf{p})$ as the unit vector that connects point $\mathbf{p}$ to its closest boundary point. One of the ways to identify skeletal points is to investigate the distance function gradient vector which is multivalued at the skeletal points. To do this investigation, we use a measure called AOF (Average Outward Flux).

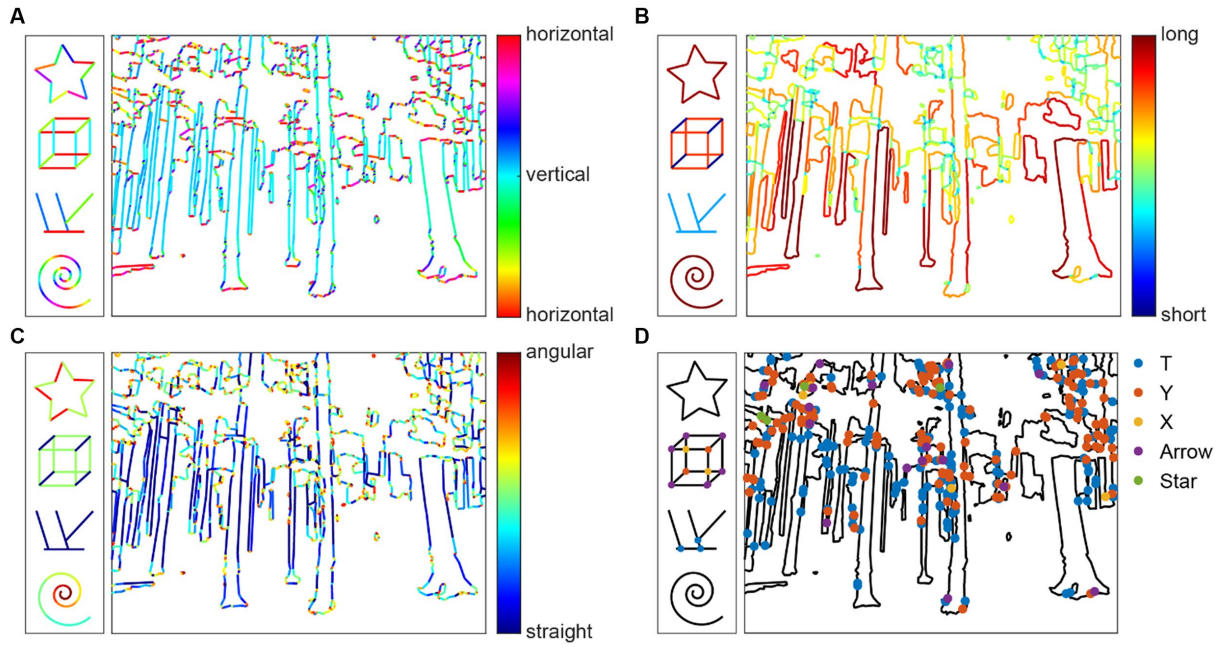


FIGURE 2
Visualization of contour properties Orientation **(A)**, Contour Length **(B)**, Curvature **(C)**, and Contour Junctions **(D)** for the example images as well as four simple test shapes.

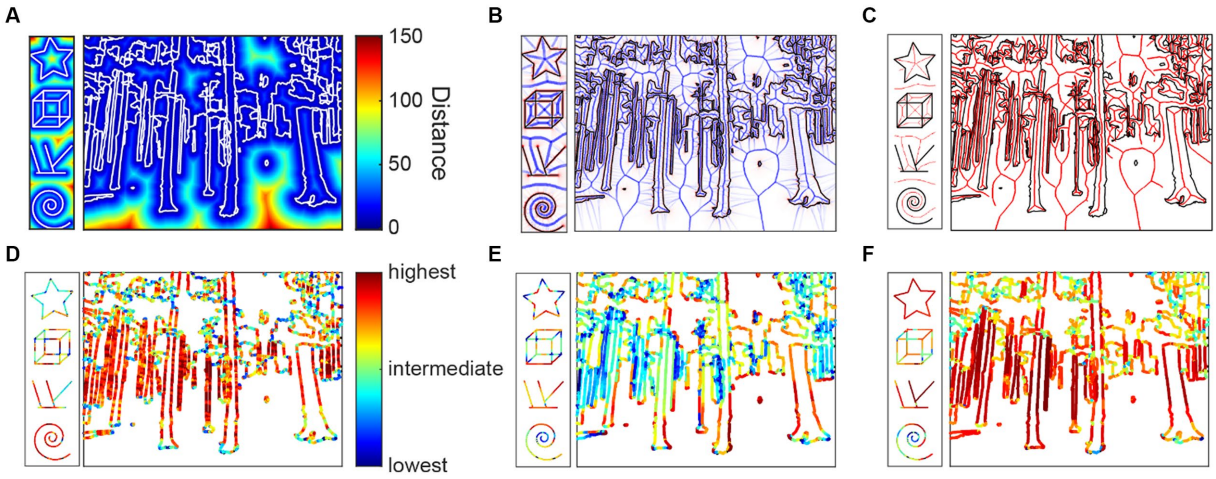


FIGURE 3
Medial axis properties. **(A)** Distance function from the contours (white); **(B)** Average outward flux (blue) from the contours (black); **(C)** Medial axes (red) between the contours; **(D)** Parallelism scores mapped onto the contours; **(E)** Separation scores; **(F)** Mirror symmetry scores.

To compute AOF, we compute the outward flux of $\dot{\mathbf{q}}$ through shrinking circular neighborhoods all over the image:

$$AOF = \frac{\int_{\partial R} \langle \dot{\mathbf{q}}, N\, \partial S \rangle}{\int_{\partial R} \partial S}$$

where ∂R is the boundary of the shrinking circular neighborhood and N is the normal to the boundary (Figure 3B). By analyzing the behavior of AOF, we can classify each point into a medial axis or non-medial axis point. In particular, any points that are not on the skeleton have a limiting AOF value of zero, so the medial axis is the set of points where their AOF values are non-zero (Figure 3C). The AOF value has a sinusoidal relationship with the object angle (the mid-angle between spoke vectors that connect a skeletal point to the closest boundary points). In the discrete space, we threshold the AOF based on a particular value, which means that we keep skeletal points that have object angles above a certain degree. The object

angle can be provided as an optional parameter for the **computeMAT** function, with a default value of 28 degrees (Rezanejad, 2020). We compute three local relational properties based on the medial axis transform.

Parallelism is computed as the local ribbon symmetry by computing the first derivative of the radial distance function along the medial axis. A small first derivative (small change) indicates that the contours on either side of the medial axis are locally parallel to the medial axis and, thereby, to each other (Figure 3D).

Separation is computed as the absolute value of the radial distance function. Separation is related to the Gestalt grouping rule of proximity (Figure 3E).

Mirror symmetry is generally understood to be the symmetry generated by reflection over a straight axis. Reflections over bent axes are not perceived as mirror symmetric. We therefore compute the curvature of the medial axis as a measure of local mirror symmetry. The straighter the medial axis is, the stronger is local mirror symmetry (Figure 3F).

These properties are initially computed on the medial axis and then projected back onto the contours of the line drawing and normalized to [0,1]. Note that not all contour pixels may receive a valid MAT property, since the projection from the medial axis back to the contour pixels is not a surjective function. As mentioned above, we apply a small threshold on the AOF values in discrete pixel space to compute a medial axis that is thin, smooth and does not cover the entire area of the interior shape. While the analytical formulation of the medial axis is a one-to-many mapping from skeletal points to the boundary that ensures that all the boundary points are reconstructable in the discrete space, we may lose a small portion of the boundary points that will not be associated with the skeletal points.

Functions: **computeMAT**, **computeMATproperty**, **mapMATtoContour**, **computeAllMATproperties**

To compute statistics over the MAT properties along the contours, the contours are mapped to the vectorized line drawing. Histograms with equally spaced bins (default: 8 bins) to cover the interval [0,1] are computed over all contour pixels with valid MAT properties.

Functions: **MATpropertiesToContours**, **getMATpropertyStats**, **computeAllMATfromVecLD**

Similar to the contour properties, the following fields are added to the vecLD struct for each MAT property:

property:	(1, numContours) cell array; each cell contains a vector of properties for the corresponding contour segments
propertyMeans:	(1, numContours) array with the means of property over each contour
property_allX:	x coordinates of all contour pixels with a property score
property_allY:	y coordinates of all contour pixels with a property score
property_allScores:	The property scores for all contour pixels
propertyHistograms:	(numContours, numBins) array with the property histograms for the individual contours
propertyNormHistograms:	the same but normalized by the total number of contour pixels
propertySumHistogram:	(1, numBins) array: the property histogram for the entire image – the sum of propertyHistograms
propertyNormSumHistogram:	the same but normalized by the total number of contour pixels
propertyBins:	(1, numBins) array: the centers of the histogram bins

MAT properties are easily visualized in a figure using **drawMATproperty** or drawn into an image using **renderMATproperty**.

The histograms for all image properties can be written into a table for a set of images for further statistical analysis. The function **allLDHistogramsToTable** generates such a table, which can then be used with Matlab's statistical functions or be written to a CSV file for further processing in R or other analysis software.

We have used these functions to investigate the role of MAT-based features for computer vision (Rezanejad et al., 2019, 2023), eye movements (Damiano et al., 2019), neural representations of symmetry (Wilder et al., 2022), to investigate esthetic appeal (Damiano et al., 2021b; Farzanfar and Walther, 2023) and image memorability (Han et al., 2023).

Splitting functions

Splitting the contours in a line drawing into two halves based on some statistical property allows for the empirical testing of the causal involvement of that property in some perceptual or cognitive function. We provide functions for separating contours into two drawings by different criteria:

splitLDbyProperties: allows for the splitting according to one image property or a combination of image properties. The function also contains an option to generate a random split of the contours. We have used this function to split images by their MAT properties for behavioral and fMRI experiments as well as for computer vision analyses (Rezanejad et al., 2019, 2023; Wilder et al., 2022).

splitLDbyHistogramWeights: allows for splitting the contours according to more fine-grained weights for the individual feature histograms.

splitLDbyStatsModel: splits the contours by the output of a statistical model, trained with the contour and MAT properties as its features. This function has been used to split contours according to predicted esthetic appeal (Farzanfar and Walther, 2023) or according to predicted memorability (Han et al., 2023).

splitLDmiddleSegmentsVsJunctions: Splits the contours into pieces near the contour junctions and middle segments. This method was used to investigate the role junctions for scene categorization (Wilder et al., 2018).

For an input vecLD struct, these functions generate two new, disjoint vecLD structs, each with approximately half of the pixels (Figure 4). Contours that cannot be uniquely assigned to one or the other drawing are omitted from both.

Other image transformations

We include some other specific manipulations of the line drawing images that have proven useful in manipulating images (Figure 5). The function **rotateLinedrawing** rotates the coordinates of all contours in the input vecLD structure by a given angle, and **applyCircularAperture** clips the contours to a circular aperture. In combination, these functions can be used to generate randomly rotated line drawings (Choo and Walther, 2016).

Randomly shifting individual contours within the image bounding box destroys the distribution of contour junctions while keeping the statistics of orientation, length, and curvature constant (Walther and Shen, 2014; Choo and Walther, 2016). This functionality is provided by **randomlyShiftContours**.

Datasets

We provide five data sets already processed as vecLD structs:

- A set of 475 images of six scene categories (beaches, cities, forests, highways, mountains, and offices). The line drawings were created by trained artists at the Lotus Hill Institute in Fudan, People's Republic of China.
- Line drawings of the 1,182 images in the International Affective Picture System (IAPS) (Lang et al., 2008), also created at the Lotus Hill Institute.
- Hand-traced drawings of 260 objects from (Snodgrass and Vanderwart, 1980).
- A set of 200 architectural scenes published in (Vartanian et al., 2013), traced automatically using **traceLineDrawingFromRGB**.
- Line drawings of the 900 images in the Open Affective Standardized Image Set (OASIS) (Kurdi et al., 2017).

We plan to add more datasets in the near future.

Conclusion

A major obstacle for research on the role of mid-level visual features in the perception of complex, real-world scenes has been the capability to measure and selectively manipulate these features in scenes. We offer the Mid-level Vision Toolbox to the research community as way to overcome this obstacle and enable future research on this topic. We include functionality for assessing a variety of low- and mid-level features based on the geometry of contours, as well as functions for generating contour line drawings from RGB images and functions for manipulating such drawings.

The easily accessible data structures and function interfaces of MLV Toolbox allow for future expansions of its functionality. For instance, symmetry relationships, limited to the nearest contours in the current implementation, could be expanded to include symmetries across larger scales. Another expansion could be the detection of another important grouping cue, closure of contours. Figure-ground segmentation cues, such as border ownership could be included in the future as well. Our group will continue to work on expanding the functionality of the toolbox, and we also invite contributions from other researchers.

Computational models of visual perception in humans and non-human primates have progressed rapidly in recent years, thanks to the advent of convolutional neural networks (Krizhevsky et al., 2012). Convolutional Neural Networks have been shown to correlate well with biological vision systems (e.g., Cadieu et al., 2014;

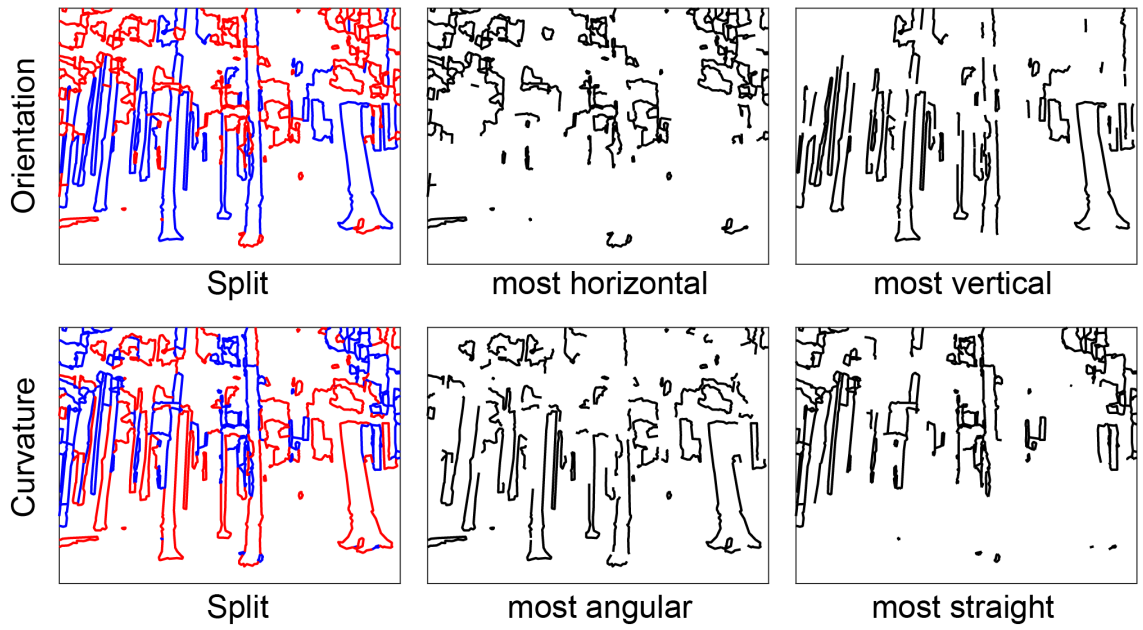


FIGURE 4
Splitting functions. Left column: Line drawings split into top (red) and bottom (blue) halves according to Orientation ("top" = horizontal, "bottom" = vertical) and Curvature ("top" = angular, "bottom" = straight). Middle column: drawings with only the top halves. Right column: drawings with only the bottom halves.

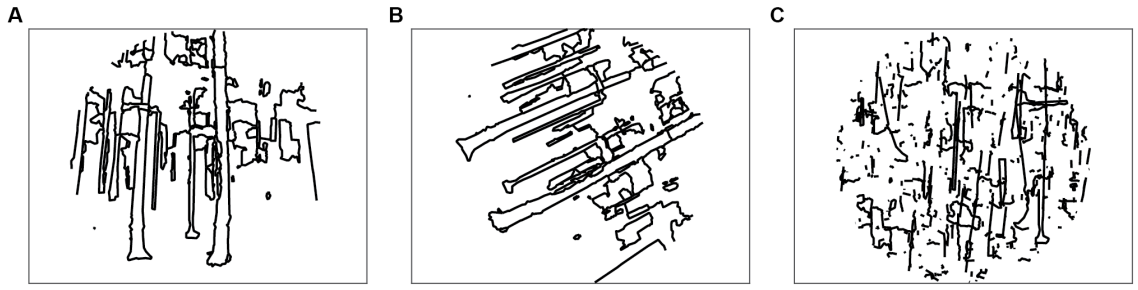


FIGURE 5
Image transformation functions. **(A)** Applying circular aperture; **(B)** rotating the image by an arbitrary angel (here: 63°); **(C)** randomly shifting contours.

Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Güçlü and van Gerven, 2015), and they are getting closer to replicating the functionality of biological systems all the time (Schrimpf et al., 2020). Why, then, should we care about hand-coded algorithms for detecting mid-level features as in the MLV Toolbox? Despite these models' increasing ability to match biological vision, the mechanisms underlying their impressive performance are barely any clearer than those underlying biological vision. We need to probe these deep neural networks empirically to better understand the mechanisms underlying the function (Bowers et al., 2022).

A century of empirical research as well as existing practice in design and architecture have unequivocally demonstrated the importance of Gestalt grouping rules for human perception (Wagemans et al., 2012) as well as esthetic appreciation (Arnheim, 1965; Leder et al., 2004; Chatterjee, 2022). To what extent convolutional neural networks learn to represent these grouping rules is an open question. We know from work with random dot patterns that the human brain represents symmetry relationships in fairly high-level areas, such as the object-sensitive lateral occipital complex (Bona et al., 2014, 2015). We are only starting to learn how the brain represents Gestalt grouping rules for perceiving complex scenes (Wilder et al., 2022).

We here provide a set of computational tools that will enable studies of the mid-level representations that arise in both biological and artificial vision systems. Although the specific computations used here to measure mid-level visual properties are unlikely to be an accurate reflection of the neural mechanisms in the human visual system, we nevertheless believe that measuring and manipulating mid-level visual cues in complex scenes will be instrumental in furthering our understanding of visual perception.

Data availability statement

The datasets presented in this study can be found at: http://mlvtoolbox.org.

Author contributions

DW and MR contributed to the algorithms and their implementation in the toolbox. DF and SH tested the code and provided suggestions and feedback for features and improvements. DW wrote the first draft of the manuscript. MR and DF wrote sections of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by an NSERC Discovery grant (RGPIN-2020-04097) and an XSeed grant from the Faculty of Applied Science and Engineering and the Faculty of Arts and Science of the University of Toronto to DW, an Alexander Graham Bell Canada Graduate Scholarship from NSERC to DF, a Connaught International Scholarship to SH, and a Faculty of Arts and Science Postdoctoral Fellowship to MR.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Arnheim, R. (1965). *Art and visual perception: A psychology of the creative eye*. Berkeley, California: Univ of California Press.

Ayzenberg, V., Kamps, F. S., Dilks, D. D., and Lourenco, S. F. (2022). Skeletal representations of shape in the human visual cortex. *Neuropsychologia* 164:108092. doi: 10.1016/j.neuropsychologia.2021.108092

Ayzenberg, V., and Lourenco, S. F. (2019). Skeletal descriptions of shape provide unique perceptual information for object recognition. *Sci. Rep.* 9:9359. doi: 10.1038/s41598-019-45268-y

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115–147. doi: 10.1037/0033-295X.94.2.115

Blum, H. (1967). *A transformation for extracting new descriptions of shape*. Cambridge, Massachusetts: MIT Press.

Bona, S., Cattaneo, Z., and Silvanto, J. (2015). The causal role of the occipital face area (OFA) and lateral occipital (LO) cortex in symmetry perception. *J. Neurosci.* 35, 731–738. doi: 10.1523/JNEUROSCI.3733-14.2015

Bona, S., Herbert, A., Toneatto, C., Silvanto, J., and Cattaneo, Z. (2014). The causal role of the lateral occipital complex in visual mirror symmetry detection and grouping: an fMRI-guided TMS study. *Cortex* 51, 46–55. doi: 10.1016/j.cortex.2013.11.004

Bowers, J. S., Malhotra, G., Duimovic, M., Montero, M. L., Tsvetkov, C., Biscione, V., et al. (2022). Deep problems with neural network models of human vision. *Behav. Brain Sci.* 1, 1–74. doi: 10.1017/S0140525X22002813

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., et al. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.* 10:e1003963. doi: 10.1371/journal.pcbi.1003963

Chatterjee, A. (2022). "An early framework for a cognitive neuroscience of visual aesthetics" in *Brain, beauty, & art*. eds. A. Chatterjee and E. R. Cardillo (Essays Bringing Neuroaesthetics in Focus, New York, NY: Oxford University Press)

Choo, H., and Walther, D. B. (2016). Contour junctions underlie neural representations of scene categories in high-level human visual cortex. *NeuroImage* 135, 32–44. doi: 10.1016/j.neuroimage.2016.04.021

Damiano, C., Walther, D. B., and Cunningham, W. A. (2021a). Contour features predict valence and threat judgements in scenes. *Sci. Rep.* 11, 1–12. doi: 10.1038/s41598-021-99044-y

Damiano, C., Wilder, J., and Walther, D. B. (2019). Mid-level feature contributions to category-specific gaze guidance. *Atten. Percept. Psychophys.* 81, 35–46. doi: 10.3758/s13414-018-1594-8

Damiano, C., Wilder, J., Zhou, E. Y., Walther, D. B., and Wagemans, J. (2021b). The role of local and global symmetry in pleasure, interest, and complexity judgments of natural scenes. *Psychol. Aesthet. Creat. Arts* 17, 322–337. doi: 10.1037/aca0000398

De Winter, J., and Wagemans, J. (2006). Segmentation of object outlines into parts: a large-scale integrative study. *Cognition* 99, 275–325. doi: 10.1016/j.cognition.2005.03.004

Desolneux, A., Moisan, L., and Morel, J.-M. (2004). "Gestalt theory and computer vision," in *Seeing, thinking and knowing: Meaning and self-organisation in visual cognition and thought*. Dordrecht: Springer Netherlands, 71–101.

Desolneux, A., Moisan, L., and Morel, J.-M., (2007). *From gestalt theory to image analysis: A probabilistic approach*. Springer Science & Business Media, New York, NY

Dollár, P., and Zitnick, C. L., (2013). Structured forests for fast edge detection, in: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 1841–1848.

Dollár, P., and Zitnick, C. L. (2014). Fast edge detection using structured forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 1558–1570. doi: 10.1109/TPAMI.2014.2377715

Elder, J. H., and Goldberg, R. M. (2002). Ecological statistics of gestalt laws for the perceptual organization of contours. *J. Vis.* 2, 5–353. doi: 10.1167/2.4.5

Elder, J., and Zucker, S. (1993). The effect of contour closure on the rapid discrimination of two-dimensional shapes. *Vis. Res.* 33, 981–991. doi: 10.1016/0042-6989(93)90080-G

Epstein, R., DeYoe, E. A., Press, D. Z., Rosen, A. C., and Kanwisher, N. (2001). Neuropsychological evidence for a topographical learning mechanism in parahippocampal cortex. *Cogn. Neuropsychol.* 18, 481–508. doi: 10.1080/02643290125929

Epstein, R., and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature* 392, 598–601. doi: 10.1038/33402

Farzanfar, D., and Walther, D. B. (2023). Changing What You Like: Modifying Contour Properties Shifts Aesthetic Valuations of Scenes. *Psychol. Sci.* doi: 10.1177/09567976231190546. [Epub ahead of print].

Feldman, J., and Singh, M. (2005). Information along contours and object boundaries. *Psychol. Rev.* 112, 243–252. doi: 10.1037/0033-295X.112.1.243

Feldman, J., and Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proc. Natl. Acad. Sci.* 103, 18014–18019. doi: 10.1073/pnas.0608811103

Field, D. J., Hayes, A., and Hess, R. F. (1993). Contour integration by the human visual system: evidence for a local "association field". *Vis. Res.* 33, 173–193. doi: 10.1016/0042-6989(93)90156-Q

Firestone, C., and Scholl, B. J. (2014). "Please tap the shape, anywhere you like" shape skeletons in human vision revealed by an exceedingly simple measure. *Psychol. Sci.* 25, 377–386. doi: 10.1177/0956797613507584

Gallant, J. L., Connor, C. E., Rakshit, S., Lewis, J. W., and Van Essen, D. C. (1996). Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *J. Neurophysiol.* 76, 2718–2739. doi: 10.1152/jn.1996.76.4.2718

Geisler, W. S., Perry, J. S., Super, B. J., and Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vis. Res.* 41, 711–724. doi: 10.1016/S0042-6989(00)00277-7

Güçlü, U., and van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* 35, 10005–10014. doi: 10.1523/JNEUROSCI.5023-14.2015

Han, S., Rezanejad, M., and Walther, D. B. (2023). Making memorability of scenes better or worse by manipulating their contour properties. *J. Vis.* 23:5494. doi: 10.1167/jov.23.9.5494

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154. doi: 10.1113/jphysiol.1962.sp006837

Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* 10:e1003915. doi: 10.1371/journal.pcbi.1003915

Koffka, K., (1935). *Principles of gestalt psychology*. Harcourt Brace and Company, New York, NY.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks" in *Advances in neural information processing systems*. eds. F. Pereira, C. J. Burges, L. Bottou and K. Q. Weinberger (United States: Curran Associates, Inc)

Kubovy, M. (1994). The perceptual organization of dot lattices. *Psychon B Rev* 1, 182–190. doi: 10.3758/bf03200772

Kurdi, B., Lozano, S., and Banaji, M. R. (2017). Introducing the open affective standardized image set (OASIS). *Behav. Res. Methods* 49, 457–470. doi: 10.3758/s13428-016-0715-3

Lang, P. J., Bradley, M. M., and Cuthbert, B. N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. University of Florida, Gainsville, FL.

Leder, H., Belke, B., Oeberst, A., and Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *Brit J Psychol* 95, 489–508. doi: 10.1348/0007126042369811

Lowe, D. G. (1987). Three-dimensional object recognition from single two-dimensional images. *Artif. Intell.* 31, 355–395. doi: 10.1016/0004-3702(87)90070-1

Lowe, D. (2012). *Perceptual organization and visual recognition*. Springer Science & Business Media. New York

Machilsen, B., Pauwels, M., and Wagemans, J. (2009). The role of vertical mirror symmetry in visual shape detection. *J. Vis.* 9:11. doi: 10.1167/9.12.11

Malcolm, G. L., Groen, I. I. A., and Baker, C. I. (2016). Making sense of real-world scenes. *Trends Cogn. Sci.* 20, 843–856. doi: 10.1016/j.tics.2016.09.003

Marr, D., (1982). *Vision: A computational investigation into the human representation and processing of visual information*.

Marr, D., and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-diemnsional shapes. *Proccedings of the Royal Soc London B.* 200, 269–294. doi: 10.1098/rspb.1978.0020

Michaelsen, E., and Meidow, J. (2019). *Hierarchical perceptual grouping for object recognition* Springer, New York.

Norcia, A. M., Candy, T. R., Pettet, M. W., Vildavski, V. Y., and Tyler, C. W. (2002). Temporal dynamics of the human response to symmetry. *J. Vis.* 2, 1–139. doi: 10.1167/2.2.1

Olshausen, B. A., and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609. doi: 10.1038/381607a0

Pasupathy, A., and Connor, C. E. (2002). Population coding of shape in area V4. *Nat. Neurosci.* 5, 1332–1338. doi: 10.1038/972

Peirce, J. W. (2015). Understanding mid-level representations in visual processing. *J. Vis.* 15:5. doi: 10.1167/15.7.5

Peterhans, E., and von der Heydt, R. (1989). Mechanisms of contour perception in monkey visual cortex. II. Contours bridging gaps. *J. Neurosci.* 9, 1749–1763. doi: 10.1523/JNEUROSCI.09-05-01749.1989

Pizlo, Z., Li, Y., and Sawada, T., (2014). *Making a machine that sees like us*. Oxford University Press, USA.

Rezanejad, M., (2020). *Medial measures for recognition, mapping and categorization*, McGill University. Canada

Rezanejad, M., Downs, G., Wilder, J., Walther, D. B., Jepson, A., Dickinson, S., et al. (2019). Scene categorization from contours: medial Axis based salience measures. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 4116–4124.

Rezanejad, M., Wilder, J., Walther, D. B., Jepson, A., Dickinson, S., and Siddiqi, K. (2023). Shape Based Measures Improve Scene Categorization. under review. *IEEE Trans. Pattern Anal. Mach. Intell.*

Sasaki, Y. (2007). Processing local signals into global patterns. *Curr. Opin. Neurobiol.* 17, 132–139. doi: 10.1016/j.conb.2007.03.003

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., and DiCarlo, J. J. (2020). Integrative benchmarking to advance Neurally mechanistic models of human intelligence. *Neuron* 108, 413–423. doi: 10.1016/j.neuron.2020.07.040

Snodgrass, J. G., and Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *J. Exp. Psychol.* 6, 174–215. doi: 10.1037/0278-7393.6.2.174

Sun, Z., and Firestone, C. (2022). Beautiful on the inside: aesthetic preferences and the skeletal complexity of shapes. *Perception* 51, 904–918. doi: 10.1177/03010066221124872

Vartanian, O., Navarrete, G., Chatterjee, A., Fich, L. B., Leder, H., Modroño, C., et al. (2013). Impact of contour on aesthetic judgments and approach-avoidance decisions in architecture. *Proc National Acad Sci* 110, 10446–10453. doi: 10.1073/pnas.1301227110

Vinje, W. E., and Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287, 1273–1276. doi: 10.1126/science.287.5456.1273

Wagemans, J. (1993). Skewed symmetry: a nonaccidental property used to perceive visual forms. *J. Exp. Psychol. Hum. Percept. Perform.* 19, 364–380. doi: 10.1037/0096-1523.19.2.364

Wagemans, J. (1997). Characteristics and models of human symmetry detection. *Trends Cogn. Sci.* 1, 346–352. doi: 10.1016/s1364-6613(97)01105-4

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., et al. (2012). A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychol. Bull.* 138, 1172–1217. doi: 10.1037/a0029333

Walther, D. B., and Shen, D. (2014). Nonaccidental properties underlie human categorization of complex natural scenes. *Psychol. Sci.* 25, 851–860. doi: 10.1177/0956797613512662

Wertheimer, M. (1922). Untersuchungen zur Lehre von der Gestalt, I: Prinzipielle Bemerkungen [Investigations in Gestalt theory: I. The general theoretical situation]. *Psychol. Forsch.* 1, 47–58. doi: 10.1007/BF00410385

Wilder, J., Dickinson, S., Jepson, A., and Walther, D. B. (2018). Spatial relationships between contours impact rapid scene classification. *J. Vis.* 18:1. doi: 10.1167/18.8.1

Wilder, J., Rezanejad, M., Dickinson, S., Siddiqi, K., Jepson, A., and Walther, D. B. (2022). Neural correlates of local parallelism during naturalistic vision. *PLoS One* 17:e0260266. doi: 10.1371/journal.pone.0260266

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 14 September 2023
DOI 10.3389/fcomp.2023.1154957

OPEN ACCESS

EDITED BY
James Elder,
York University, Canada

REVIEWED BY
Daniel Coates,
University of Houston, United States
Dejan Todorović,
University of Belgrade, Serbia
Najib Majaj,
New York University, United States

*CORRESPONDENCE
Oh-Hyeon Choung
✉ ohhyeon.choung@gmail.com

RECEIVED 31 January 2023
ACCEPTED 28 August 2023
PUBLISHED 14 September 2023

CITATION
Choung OH, Rashal E, Kunchulia M and Herzog MH (2023) Specific Gestalt principles cannot explain (un)crowding. *Front. Comput. Sci.* 5:1154957. doi: 10.3389/fcomp.2023.1154957

Specific Gestalt principles cannot explain (un)crowding

Oh-Hyeon Choung[1]*, Einat Rashal[1,2], Marina Kunchulia[3,4] and Michael H. Herzog[1]

[1]Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, [2]School of Psychology, Keele University, Staffordshire, United Kingdom, [3]Vision Research Laboratory, Beritashvili Centre of Experimental Biomedicine, Tbilisi, Georgia, [4]Institute of Cognitive Neurosciences, Free University of Tbilisi, Tbilisi, Georgia

The standard physiological model has serious problems accounting for many aspects of vision, particularly when stimulus configurations become slightly more complex than the ones classically used, e.g., configurations of Gabors rather than only one or a few Gabors. For example, as shown in many publications, crowding cannot be explained with most models crafted in the spirit of the physiological approach. In crowding, a target is neighbored by flanking elements, which impair target discrimination. However, when more flankers are added, performance can improve for certain flanker configurations (uncrowding), which cannot be explained by classic models. As was shown, aspects of perceptual organization play a crucial role in uncrowding. For this reason, we tested here whether known principles of perceptual organization can explain crowding and uncrowding. The answer is negative. As shown with subjective tests, whereas grouping is indeed key in uncrowding, the four Gestalt principles examined here did not provide a clear explanation to this effect, as variability in performance was found between and within categories of configurations. We discuss the philosophical foundations of both the physiological and the classic Gestalt approaches and sketch a way to a happy marriage between the two.

KEYWORDS

visual perception, perceptual organization, Gestalt principle, crowding, grouping, vision model, recurrent processing

Introduction

Vision has been a mystery since ancient times. Intuitively, perception seems to give us ground truth about the outer world and its objects. Based on this intuition, direct realism proposes a one-to-one mapping (bijection) between the objects of the external world and our mental representations. When there is an apple in the external world, we perceive an apple, and when we perceive an apple, there is an apple in front of our eyes (leaving dreams and mental imagery aside). However, perception can hardly be direct. For example, according to the laws of optics, objects are projected upside-down (and left-right inverted) on the retinal image, but we perceive them upright, so there must be a second transformation undoing the laws of optics, for example, when we want to grasp these objects. Philosophically speaking, perception is not direct but indirect. Still, perception may give us ground truth about the external world, at least approximately. However, the situation is worse. We see many illusory things that are not out there. For example, we see blue spiral lines in the Munker-White illusion, which are simply not there (see Figure 2 of Herzog, 2022). In this case, a simple re-transformation cannot help.

For similar reasons, philosophers such as Berkeley, Kant, and Fichte, have abandoned reasoning about the external world. Ground truth is found in these approaches in the

percepts and thoughts themselves. If I have the experience of an apple, there might be no corresponding apple out there, but my percept is undeniably true. Thus, focusing on the laws of perception and cognition is a first step toward a philosophy free of ontological commitments about an external world to which we have no direct access. The Gestaltists have largely subscribed to this continental approach of epistemology, with introspection being both the conceptual and methodological starting point. For many Gestaltists, there is more in the mind than in the world. There is information that goes "beyond the information given" (Kanizsa, 1979). For example, we clearly perceive a cross in Figure 1 (upper panel) even though there are only squares and disks.

Gestalt theory disappeared as quickly as it arose and was replaced by the physiological approach subscribing (implicitly) to indirect realism. This approach has dominated vision science for almost a century, aiming for a causal theory of perception. Physiology systematically studies how the presentation of an object affects neural responses, starting with phototransduction at the retina and continuing up the hierarchy of brain processing. One crucial aspect is that perception is genuinely ill-posed. The light that arrives at the retina (luminance) is always the product of the light shining on an object (illuminance) and the material properties of that object (reflectance). Hence, there are infinitely many possibilities for how a given luminance may have occurred (e.g., white light on a red tomato leads to the same luminance as red light shining on a white tomato). To perceive the true object properties, one needs to reconstruct the object. Because solving the ill-posed problems is mathematically not fully possible, this reconstruction may fail. Illusions and alike are rather evidences for the physiological approach than challenges.

Whereas the physiological approach has made great progress in explaining the first steps of vision (retina, LGN, V1), the processing of subsequent stages has turned out to be less straightforward. One reason may simply be that perception is not as one-to-one as assumed, i.e., perception is not only indirect, but percepts do not systematically match the objects of the external world, as in the case of the cross of Figure 1.

Predictions of the standard physiological model of perception fail, also in many classic psychophysical paradigms. Crowding is one example. In crowding, perception of a target strongly deteriorates when presented within clutter (Figure 1, lower panel, a). Crowding is the usual situation in daily life since elements are rarely presented alone (Weymouth, 1958; Bouma and Andriessen, 1968; Bouma, 1970, 1973; Strasburger et al., 1991; Levi, 2008). Crowding is traditionally explained by feature pooling or averaging (e.g., Parkes et al., 2001; Solomon et al., 2004; Pelli, 2008; Greenwood et al., 2009, 2017; Dakin et al., 2010; Rosenholtz et al., 2012). Whereas, pooling can well explain results with simple stimuli, it fails as soon as stimulus configurations become slightly more complex (Figure 1, lower panel).

For example, vernier offset discrimination drops drastically when the vernier is presented within a square well in line with pooling and other low-level physiological explanations. However, adding more squares improves performance almost to the level of the no crowding condition (Figure 1, lower panel; Manassi et al., 2012, 2013, 2015, 2016; Herzog and Manassi, 2015; Herzog et al., 2015, 2016; Choung et al., 2021). We argued that the Vernier information is recovered because the target and the squares are in different perceptual groups (Figure 1, lower panel, b and d), which is not the case when only one square is presented (Figure 1, lower panel, a). We call this release from crowding "uncrowding," even when performance in the uncrowding condition does not reach the performance level in the no crowding, baseline condition. These results are not restricted to crowding and vernier stimuli but occur all over the place in vision as well as in audition and haptics (e.g., peripheral vision: Bouma and Andriessen, 1968; Toet and Levi, 1992; Chung et al., 2001; foveal vision: Flom et al., 1963; Danilova and Bondarko, 2007; Lev et al., 2014; Coates et al., 2018; verniers: Malania et al., 2007; Saarela and Herzog, 2008, 2009; Sayim et al., 2008, 2010, 2011, 2014; Saarela et al., 2009, 2010; Gabors: Chicherov et al., 2014; Chicherov and Herzog, 2015; Jastrzębowska et al., 2021; audition: Oberfeld and Stahn, 2012; touch: Overvliet and Sayim, 2016). Thus, we are back to square one, i.e., the Gestalt times.

Here, we asked whether Gestalt principles can do better than explanations from the physiological approach. Gestalt principles have been studied over centuries and are considered fundamental of perceptual organization (von Ehrenfels, 1890; Wertheimer, 1912, 1922, 1923; Köhler, 1920; Koffka, 1935; Metzger, 1936; Metzger et al., 2006; reviews: Todorović, 2007; Wagemans et al., 2012a,b). Gestalt principles include symmetry, proximity, similarity, common fate, good continuation, closure, parallelism, synchrony, common region, element, and uniform connectedness. In this study, we applied four such principles that pertain to the structure of the configuration (rather than the isolated principle), namely, symmetry, good continuation, closure, and repetition. Note that while the classic displays used by the earlier Gestaltists depicted specific instances of these principles, modern studies have offered more examples that are easier to apply in complex configurations such as the ones employed in our study (symmetry: Sasaki et al., 2005; good continuation: e.g., Lezama et al., 2016; closure: e.g., Han et al., 1999; repetition: Treder and van der Helm, 2007; van der Helm, 2014). Specifically, our displays depicted configurations of stars and squares, assuming grouping by shape similarity occurs in all of them. Our grouping manipulation, then, concerned other principles that were imposed on the similarity grouping (see Figure 1). The rationale of the following experiments is that uncrowding should occur when the central square is grouped with other squares. Consequently, we hypothesized that Gestalt principles could explain (un)crowding, as crowding would affect performance in a similar manner in configurations that employed the same Gestalt principle. Moreover, we tried to further categorize our configurations as more nuanced instances (e.g., symmetry with 1 or 2 axes), to potentially uncover more subtle effects of these principles on (un)crowding. However, this was not what we found. Our results only showed symmetry to have a minor advantage in our study, with no other systematic difference in performance.

Materials and methods

Participants

Thirty-one participants took part in the experiments. Eleven out of the 31 participants were excluded after a calibration session

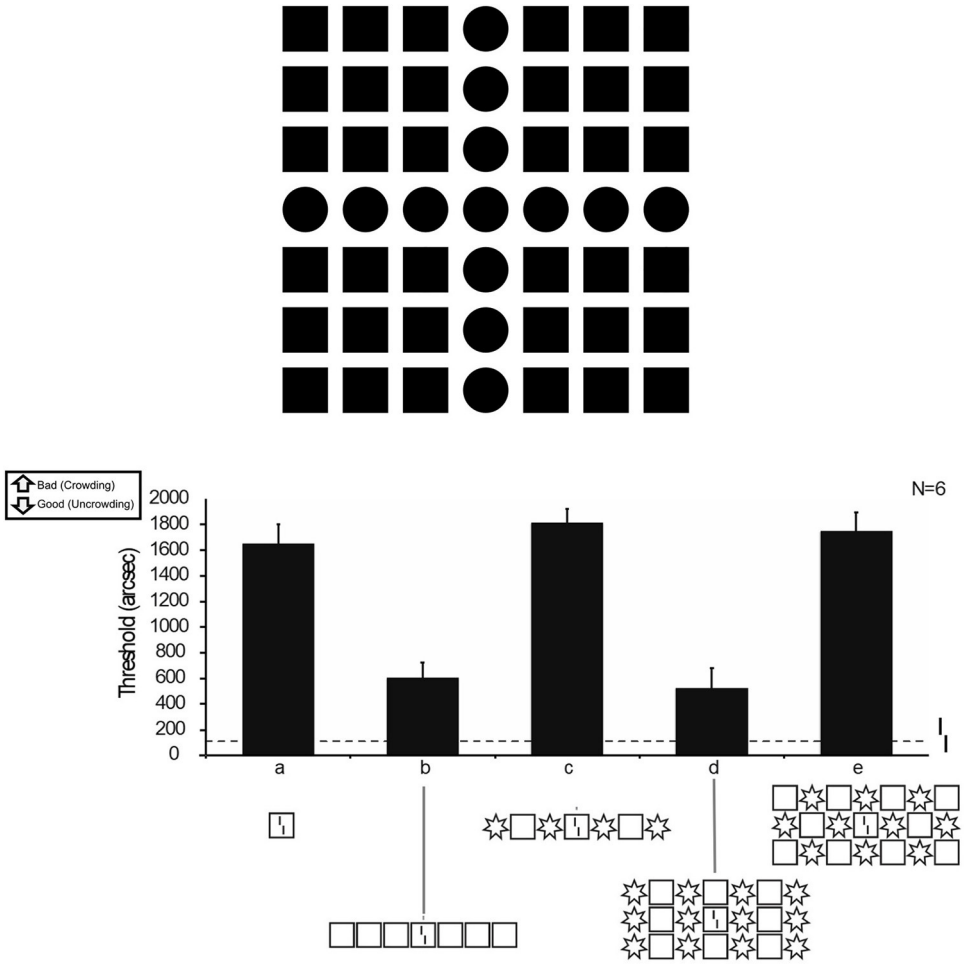


FIGURE 1
(Upper panel) We see a cross even though there are only squares and disks. **(Lower panel)** Classic models of vision fail to explain crowding in complex configurations. The x-axis shows different configurations. The y-axis shows the corresponding vernier threshold. A high value represents poor performance, a lower value represents good performance. The dashed line shows the performance of the vernier alone condition. When one square (a) is presented surrounding the vernier, performance deteriorates, i.e., crowding. When presenting 7 squares (b), performance improves drastically, i.e., uncrowding. When presenting squares and stars with different configurations (c–e), performances differ depending on the configuration. Note that the local configuration of all the conditions (a–e) is identical, i.e., a square surrounding a vernier. With permission, figure modified from Manassi et al. (2016).

because they did not show strong crowding in the basic one-square condition, which is a prerequisite for a release of crowding (see *Calibration session*). Hence, we retained the data of 20 participants (mean age: 21.6 ± 1.6, 10 females, all right-handed, 7 with left eye dominance). All participants had normal or corrected to normal visual acuity in the Freiburg Visual Acuity Test, as indicated by a binocular score greater than 1.0 (Bach, 1996). Observers gave written consent before the experiments. All experiments were conducted in accordance with the Declaration of Helsinki (World Medical Association, 2013), except for preregistration, and were approved by the local ethics committee (Beritashvili Centre of Experimental Biomedicine, Georgia).

Apparatus

Stimuli were displayed on a gamma-calibrated 24-inch ASUS VG248QE LCD monitor (1,920 x 1,080 px, 120 Hz). The room was dimly illuminated (~0.5 lux). The viewing distance was 75 cm, and the participant's chin and forehead were positioned on a chin rest. Responses were collected using wireless hand-held push buttons. In the Vernier discrimination task, when no response was registered within 3 s, the trial was repeated randomly within the same block. A feedback tone was given for incorrect responses (high tone, 600 Hz) and omissions (low tone, 300 Hz).

General procedures

Three tasks (Figure 2) were carried out with 40 flanker configurations (Figure 3). The three tasks were a vernier discrimination task (VCrowd), a vernier standout ranking task (VRank), and a rating task (Rate). The VRank and the Rate tasks were tested twice. The experiment was conducted on 5 days within 2 weeks (day 1–3: calibration and VCrowd, day 4: VRank twice and Rate, day 5: Rate). Before the first experimental

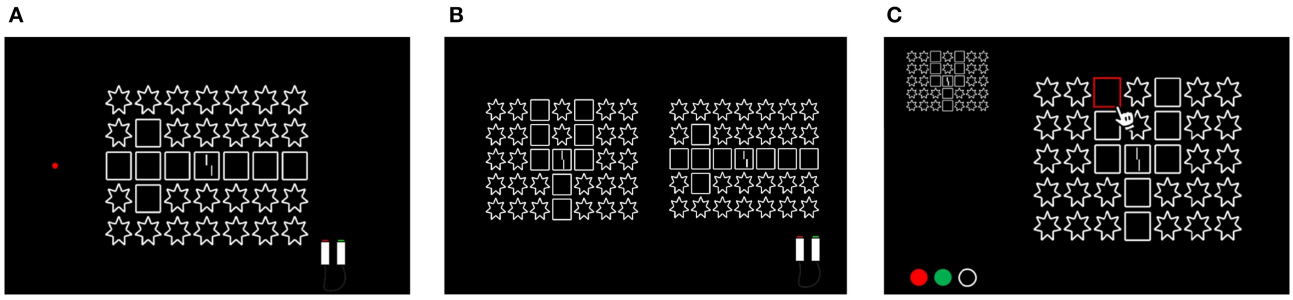


FIGURE 2
(A) VCrowd task: The task was to discriminate whether the lower Vernier bar is offset to the left or right compared to the upper bar. **(B)** VRank task: Vernier standout ranking task. Two stimuli were presented side-by-side with the same size. The task was to choose in which flanker configuration the Vernier target stands out more strongly. All possible pairwise comparisons, i.e., 40 x 39/2 pairs of configurations, were tested. **(C)** Rating task: Participants were asked to rate how much the vernier stands out (i) from the other elements, (ii) how much the center group stands out from the other elements, and (iii) how strongly the elements of the center group are grouped with each other.

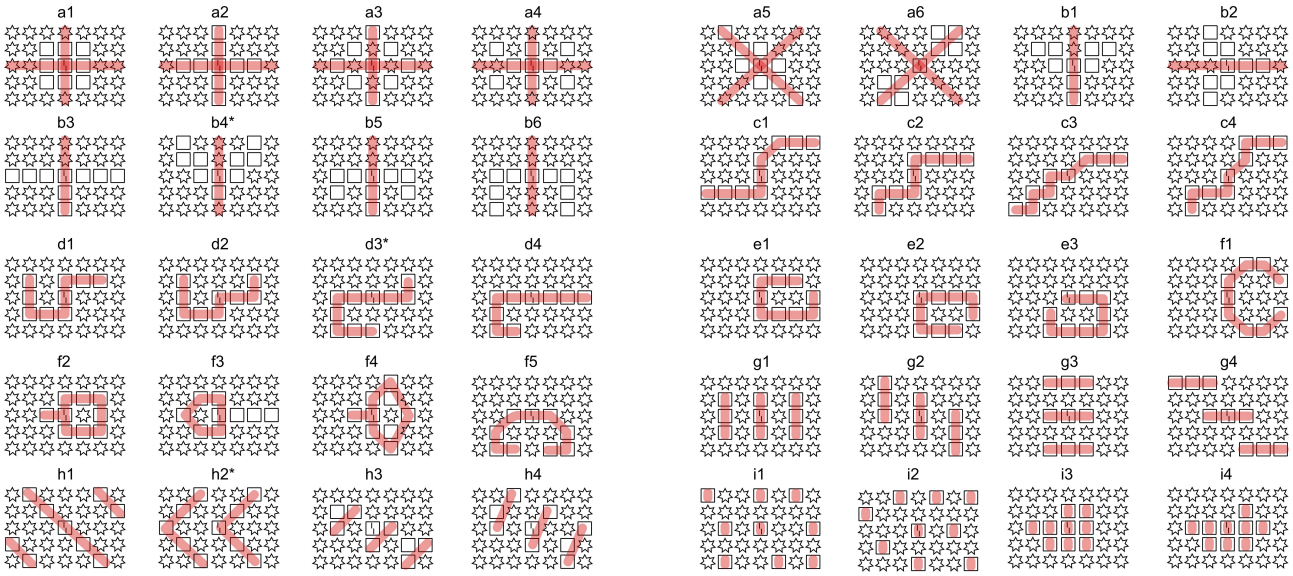


FIGURE 3
Flanker configurations. Red lines indicate the tested Gestalt principle; these lines are for illustration purposes only and were not presented during the experiment; Symmetry- a1-b6 [a1-a6- symmetry with 2 axis (Symm2); b1-b6 – symmetry with 1 axis (Symm1)]; good continuation- c1-d4 [c1-c4 – stretched (contStret); d1-d4-curled (ContCurl)]; Closure- e1-f5 [e1-e3-closure only (Close); f1-f5-closure with symmetry (CloseSymm)]; repetition- g1-h4 [g1-g4- repetition on cardinal axes (Repeat); h1-h4-repetition diagonal (RepeatDia)]; random- i1-i4 [i1andi2-random spaced (RandSpace) and random condense (i3&i4: RandCond) group]. Note that the RandCond configurations could also be considered as grouping by proximity, which is another grouping principle. Most of the configurations were composed of 9 squares and 26 stars (*indicates three configurations, which had 10 squares and 25 stars, b6, d3, and h2). Therefore, low-level features, such as pixel values, were roughly the same across configurations.

session, all participants went through a calibration session to adjust flanker-target distance individually.

Stimuli

Stimuli were white (100 cd/m^2), presented on a black background with a luminance below 0.3 cd/m^2. Participants were asked to fixate on a red fixation dot (diameter = 8 arcmin, 20 cd/m^2). Each stimulus was composed of a Vernier target, flanking squares and stars. The Vernier target was composed of two 40 arcmin long, 1.8 arcmin wide vertical bars. The gap between the two bars was 4 arcmin. Left/right offsets were balanced within a block. The Vernier target was surrounded by 35 flanker elements, which were mostly composed of 9 squares and 26 stars, except for 3 configurations which contained 10 squares and 25 stars. Squares and stars were positioned in 5 rows and 7 columns, as in Figure 3, and there were 40 different configurations. Each flanker configuration followed one of four Gestalt principles; symmetry (in 1 or 2 axes), good continuation (stretched or curled), closure (only or with symmetry), repetition (on cardinal axis or on diagonal axis), or were chosen to not include any obvious grouping principle. The central flanker was always a square, and the Vernier target was always located within this square. Except for the VCrowd task, each square was composed of four 120 arcmin long lines, and each star was composed of seven 48 arcmin long lines. The center-to-center

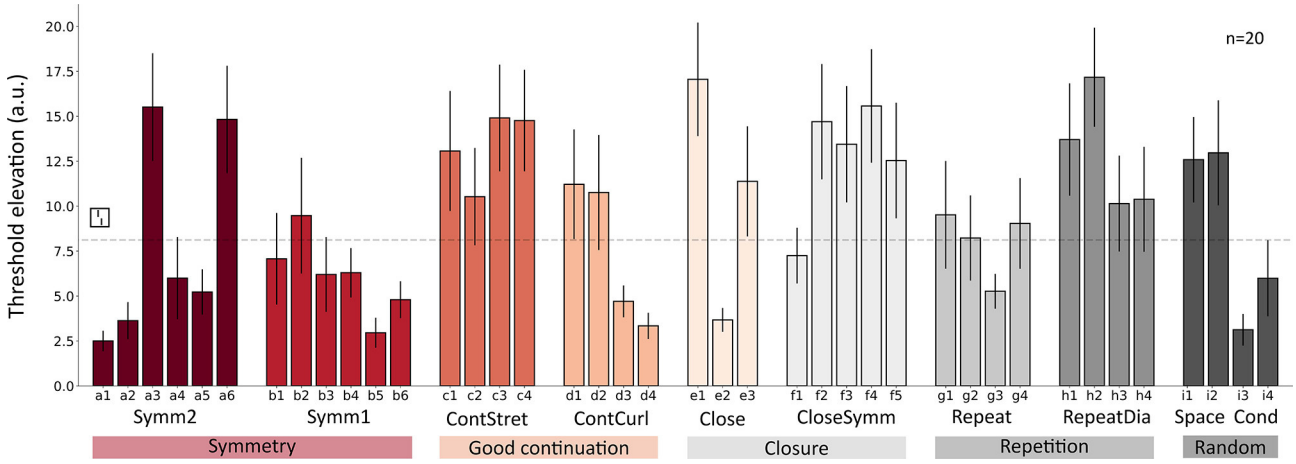


FIGURE 4
Performance for each configuration. Each color represents a different Gestalt principle. Configurations are presented in the same order as in Figure 3. The y-axis shows threshold elevation compared to the vernier only condition. Mean ± SEM. The hatched line shows performance in the one-square reference condition.

distance between flankers was 150 arcmin. For the VCrowd task, the square and star sizes and the gap between flankers were individually adjusted in a calibration session (details in *Calibration session*). The side length of the squares was 84–114 arcmin, and the gap between squares was 21–28.5 arcmin depending on observers.

Each configuration was presented at the center of the screen, and the fixation dot was presented at an eccentricity of 9 degrees to the left. Hence, stimuli were presented at 9 degrees in the visual periphery. The chin-rest was placed 75 cm from the fixation dot. Psychophysics Toolbox was used to present the stimuli (Brainard, 1997; Pelli and Vision, 1997; Kleiner et al., 2007). To avoid visual aftereffects, a small spatial jitter was applied to the entire stimulus within a 3 pixels range from trial to trial.

Procedures

Calibration session

To avoid floor and ceiling effects, each participant went through a calibration session before the main experiment. The calibration session was composed of two conditions. First, 1 or 2 blocks with the Vernier alone condition (160 trial per block) were tested to familiarize observers with the Vernier task (only participants with thresholds larger than 200 arcsec were tested in the 2nd block). Second, up to 7 blocks with a vernier surrounded by one square (80 trial/block) were tested to find the spatial parameters that produce strong crowding and, thus, allow for a release from crowding, i.e., uncrowding. We reduced the flanker size and the flanker-to-flanker distance gradually, until the threshold of the one-square condition reached at least 6 times the threshold of the Vernier alone condition. We excluded participants whose thresholds were still below this criterion even after reducing the square size to 70%. In total, 11 of 31 participants were excluded. For the remaining 20 participants, the mean threshold for the vernier alone condition was 142.30 ± 45.48 and 935.84 ± 188.53 for the one square condition. Note that crowding effects existed in the 11 excluded participants as well, but not to the extent we requested, which is at least 6 times the threshold in the one-square condition. We had this high threshold to make sure that a missing release of crowding is a clear indication of a null result.

VCrowd task

The vernier discrimination task (Figure 2A), the stimulus (Vernier + flankers) was presented for 150 ms in the center of the screen, and participants were asked to discriminate whether the lower bar was offset either to the left or right compared to the upper bar, by pressing the left or right button, respectively. Each configuration was tested in a block of 100 trials. The vernier target without flankers was presented in the first trial of each block to reduce target-location uncertainty. We used the PEST (Parameter Estimation by Sequential Testing) stair-case procedure (Taylor and Creelman, 1967) to determine testing levels (offsets). The PEST procedure changes the test levels depending on the recent history step-wise. Test levels are only changed when the hit rate is above or below the threshold criterion of 75%. The procedure ended after 100 trials, and a threshold (*Thresh*) was derived from *post-hoc* fitting of a psychometric function to the data (details in *Data analysis*).

VRank task

The Vernier standout ranking task (Figure 2B), two flanker configurations were presented simultaneously side-by-side, and participants were asked to choose from which flanker configuration the Vernier target stands out more strongly, i.e., a "win" (Figure 2B). The stimulus was presented with unlimited time. Overall, 718 (20*39) pairs of configurations were tested twice. The responses from the two identical comparisons were averaged. We ranked the order of the configurations from 1 to 40, by counting the number of "wins" in each pair of comparisons. When two or more configurations had the same number of "wins," the winner is the winner in the direct comparison between the configurations. In addition to the individual Rank order per participant, a global rank

(GlobRank) was obtained by using a similar process, by pooling the number of "wins" from the 20 participants' responses.

Rate task

The rating tasks (Figure 2C). As in the VCrowd task, the stimulus was presented for 150 ms. Four questions were asked. First, participants rated how much the vernier target stands out from the rest of the configuration on a scale from 1 to 5 (*VStandRate*). Second, the stimulus was presented with unlimited time, and the participants were asked to assign each flanker element to different sub-groups. Then, the observers were asked to rate on a scale from 1 to 5, first, how much each sub-group stands out from the other groups (*GStandRate*), and second, how strongly the elements in each group grouped together (GGroupRate)?

Hence, we determined five measures: crowding threshold (Thresh; from the VCrowd task), global vernier standout ranking (GlobRank; from the VRank task), vernier standout rating (VStandRate; from the Rate task), group standout rating (GStandRate; from the Rate task), and grouping strength (GGroupRate from Rate task).

Data analysis

We fitted a cumulative Gaussian function to the data and determined the vernier offset threshold (Thresh), for which 75% of correct responses were reached. High thresholds indicate inferior performance, and low thresholds indicate good performance. The Psignifit 2.5 toolbox (Fründ et al., 2011) was used for psychometric function fitting. We computed threshold *elevation* for each condition and each observer, i.e., we divided the threshold in each condition by the threshold in the Vernier alone condition. Data were log-transformed to bring the data closer to normality. No obvious violation was detected by visual inspection.

Using R (R Core Team, 2019) and *lme4* package (Bates et al., 2015), we computed linear mixed-effects models (LMM) to account for random variations due to individual differences. The fixed and random effects are specified for each experiment. The model significance (p-value) was obtained through likelihood ratio tests (χ^2) by comparing nested models. For each fitted model, using *MuMIn* package (Barton, 2020), we computed the effect size (r^2), i.e. the explained variance, when including (conditional r_c^2) and excluding (marginal r_m^2) the random effects (Nakagawa and Schielzeth, 2013; Johnson, 2014; Nakagawa et al., 2017). Posthoc multiple comparisons of means were computed with *multcomp* package (Hothorn et al., 2008).

Intra-rater reliability for the Rate task was carried out by using ordinal alpha (Zumbo et al., 2007) to account for ordinality of the measures (VStandRate, GStandRate, and GGroupRate). The psych package was used (Revelle, 2021).

Correlations between the measures were computed using Spearman rank correlation (Spearman, 1904), as four measures among five were in ordinal scale. Moreover, to account for the individual variances and potential violation of normality of the data, the significance of the correlations was obtained through randomization tests (details in Supplementary Method Section; Mohr and Marcon, 2005; Bakdash and Marusich, 2017).

Results

Intra-rater reliability

We computed ordinal alpha (Zumbo et al., 2007; Gadermann et al., 2012) to test intra-rater reliability for the three measures (VStandRate, GStandRate, GGroupRate) of the Rate Task and found good reliability for most configurations having alphas larger than 0.7 (Cohen, 1988; McHugh, 2012): VStandRate: $\alpha \in [0.730, 0.992]$; GStandRate: $\alpha \in [0.708, 1]$; GGroupRate: $\alpha \in [0.595, 1]$, except for two configurations for the GGroupRate. For this reason, we used the averaged rating values in the subsequent analyses.

Gestalt principles cannot explain (un)crowding

Here, we tested to what extent perceptual grouping can be explained by the Gestalt principles used here, and whether certain principles contribute more strongly than others. For example, flanker configurations with 2 symmetry axes should lead to good performance, i.e., less crowding, whereas we expected poor performance, i.e., strong crowding, for irregular configurations. We tested 40 configurations, which followed five different Gestalt principles.

Performance was hardly explained by Gestalt principles. Figure 4 shows the crowding strength (Thresh) for each configuration. Importantly, crowding levels related to the same Gestalt principle were not consistent. For example, four configurations with two symmetry axes showed uncrowding (red bars' values smaller than that of the gray dotted line; Figure 4 a1, a2, a4, and a5), whereas the other two showed strong crowding (red bars' values larger than that of the gray dotted line; Figure 4 a3 and a6). We used a linear mixed effect model (LMM) with the fixed effect of Gestalt principles and random intercepts of configurations and participants. The fixed effect was significant (likelihood ratio test between models including and excluding the fixed effect: $\chi^2(4) = 14.352$, $p < 0.01$). However, *post-hoc* Tukey's HSD comparison showed that no Gestalt principle explains the data better than other ones in general, except that configurations with symmetry had better performance than that with closure (details in Supplementary Table 2). In addition, we wondered whether the performances between the configurations sharing the same Gestalt principle correlate with each other. As shown in Supplementary Figure 3, performances within the same Gestalt principles (Supplementary Figure 3, inside red dotted lines) did not have higher correlations than those from different principles (Supplementary Figure 3, outside red dotted lines).

Subjective grouping and segmentation measures are correlated with crowding level but not with a specific principle

Thus, why do Gestalt principles not explain the performance in the VCrowd task? Two options come to mind: (1) Gestalt principles

10.3389/fcomp.2023.1154957

TABLE 1 Absolute values of correlation coefficients, significance, and 95% confidence interval (CI).

Comparisons	*Coeff.* r_{mean}	Significance (p_{Bonf})	95% *CI*
Thresh-GlobRank	$r_{mean} = 0.44$	$p < 0.001$	95% $CI = [0.38, 0.49]$
Thresh-VStandRate	$r_{mean} = 0.17$	$p < 0.01$	95% $CI = [0.10, 0.23]$
Thresh-GStandRate	$r_{mean} = 0.18$	$p < 0.01$	95% $CI = [0.11, 0.24]$
Thresh-GGroupRate	$r_{mean} = 0.14$	$p < 0.01$	95% $CI = [0.07, 0.21]$
GlobRank-VStandRate	$r_{mean} = 0.33$	$p < 0.001$	95% $CI = [0.35, 0.38]$
GlobRank-GStandRate	$r_{mean} = 0.3$	$p < 0.001$	95% $CI = [0.23, 0.36]$
GlobRank-GGroupRate	$r_{mean} = 0.33$	$p < 0.001$	95% $CI = [0.26, 0.39]$
VStandRate-GStandRate	$r_{mean} = 0.24$	$p < 0.001$	95% $CI = [0.17, 0.30]$
VStandRate-GGroupRate	$r_{mean} = 0.19$	$p < 0.001$	95% $CI = [0.13, 0.25]$
GStandRate-GGroupRate	$r_{mean} = 0.5$	$p < 0.001$	95% $CI = [0.44, 0.55]$

are not the major driver of grouping or (2) (un)crowding is not mediated by grouping.

First, we used LMMs to test if the Gestalt principles are a predictor for the grouping and segmentation measures (Rank, VStandRate, GStandRate, GGroupRate). An LMM with a fixed effect of Gestalt principles and random intercepts of configurations and participants was computed for each measure. Most of the models showed a significant fixed effect, except for VStandRate (GlobRank: $\chi^2(4) = 19.969$, $p_{Bonf} < 0.01$; VStandRate: $\chi^2(4) = 7.406$, $p_{Bonf} = 0.464$; GStandRate: $\chi^2(4) = 18.662$, $p_{Bonf} < 0.001$; GGroupRate: $\chi^2(4) = 14.632$, $p_{Bonf} < 0.05$; detailed estimates in Supplementary Table 4). However, similar to the previous experiment, no single Gestalt principle had high rates or low rates in general, except the symmetry configurations showed better ratings than other principles (*post-hoc* Tukey's HSD test; GlobRank: symmetry vs. closure $p < 0.001$; symmetry vs. contiunous $p < 0.05$; symmetry v.s. repetition $p < 0.05$; GStandRate: symmetry vs. closure $p < 0.01$, symmetry vs. random $p < 0.001$; GGropRate: symmetry vs. closure $p < 0.05$, symmetry v.s. random $p < 0.01$; details in Supplementary Table 5).

Next, we tested correlations between the performance measure (Thresh) and the grouping and segmentation measures. As expected, all the measures had significant correlations, even after Bonferroni correction (details in Table 1). We computed Spearman's Rank correlation to account for the ordinal scales; significance was obtained by randomization tests (details in Supplementary material). Figure 5 shows the average of absolute Spearman *r* coefficients. The full results for each configuration and the distributions of the randomization test are presented in Supplementary Figure 1. The correlation between (un)crowding (Thresh) and Vernier standout (GlobRank) measures was high; two Vernier standout measures had a strong correlation (GlobRank-VStandRate).

Altogether, these results indicate that subjective ratings of grouping and segmentation are indeed highly correlated with the (un)crowding performance. However, grouping processes could not be explained by classic Gestalt principles.

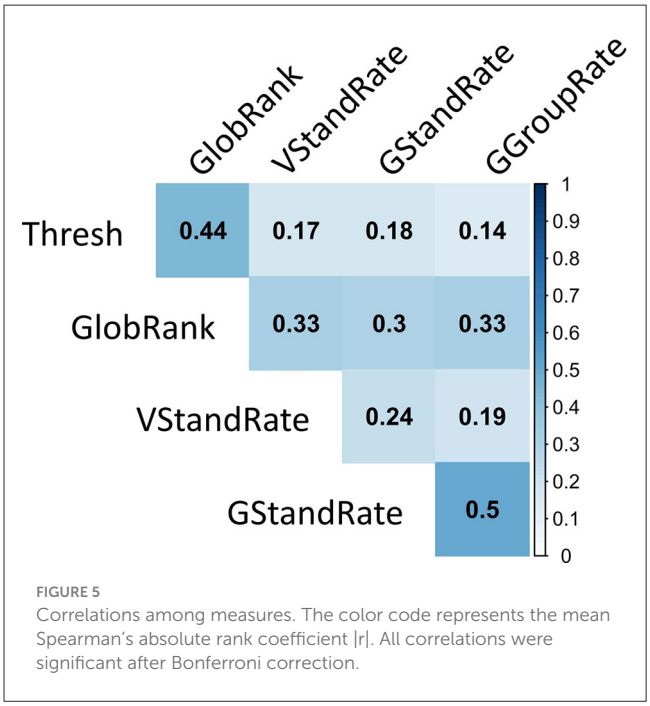


FIGURE 5
Correlations among measures. The color code represents the mean Spearman's absolute rank coefficient |r|. All correlations were significant after Bonferroni correction.

Low-level factors

The correlation between the subjective ratings and the offset discrimination task suggests that higher-level grouping is crucial. To further support this claim, we show the number of squares neighboring the central square, a high-level feature, shows higher correlations with performance than the number of white pixels, a corresponding low-level feature.

Figure 6 shows the correlations between the mean performances across participants and model predictions. Correlations between threshold elevations and the number of connected squares, discounted by distance, show a strong correlation (r_{square} (38) $= -0.60$, $CI_{95\%} = [-0.75, -0.33]$, $p < 0.001$). However, flanker pixel values, regardless of the local crowding window restriction, show poor correlation (r_{pixel} (38) $= -0.03$, $CI_{95\%} = [-0.35, 0.29]$, $p = 0.87$).

We analyzed the predictability of the two models using two methods. First, we used LMMs, which had each of the model estimates as the fixed effects. We found that the number of connected squares has a significant effect on thresholds, unlike the number of pixels. For each LMM, the fixed effect was model estimates for each configuration, and each participant was considered a random intercept. There were significant fixed effects for the number of directly connected squares based models, but not for the pixel value based models (details in Supplementary Table 1). Although the effects could only explain 6.0 % of the variances (r^2_m, square model; for the other models, see Supplementary Table 1), it was still better than the pixel estimators (0.0%, pixel model). Note that explained variances, including the random intercept across all the models, were comparable, 40%−45% (r^2_c).

Next, we tested predictability with the leave-one-out cross-validation (LOOCV) method. Here, we validated the explained variance of each participant's performance from other participants' performances. We fitted the model estimates of threshold elevation of 19 participants. We obtained an r^2-value (explained variance) by using data that was not included in model estimation. We repeated the computation 20 times (for each participant), then we averaged the r-squared values from 20 iterations to get the final explained variance of each model. As a result, similar to LMMs' estimates, the number of directed squares discounted by their distances predicted the crowding level partially ($r^2_{\mathrm{LOOCV-square}} = 0.164$), whereas pixel values did not ($r^2_{\mathrm{LOOCV-pixel}} = 0.015$).

These results indicate that none of the models can truly explain crowding and uncrowding. There were large performance variances across participants and across configurations. However, the number of directly connected squares and the remaining flankers' distances partly captured uncrowding. For full analyses of variations of these two models, see Supplementary Models Section.

Discussion

(Un)crowding is ubiquitous. Still, there is no consensus about the underlying mechanisms. Classic explanations, such as pooling, fail to explain (un)crowding. As shown here and previously, the stimulus configuration across more or less the entire visual field matters. For example, the number of squares and stars is identical in almost all configurations in the experiments above, but performance varies strongly even though all configurations contain the central square. In addition, the size of all configurations is 17.5 deg in the horizontal and 12.5 deg arcmin in the vertical direction, spanning a large part of the visual field. Thus, the specific configuration across a large part of the visual field matters.

We proposed that the stimulus configuration is parsed into different groups and crowding occurs, if at all, only within a group (Herzog et al., 2016). Hence, grouping is key in crowding. Here, we asked whether specific Gestalt principles, aimed to explain grouping, can explain crowding and uncrowding (in this respect, crowding could have been an objective test for Gestalt processing replacing the subjective reports usually used in the field). However, we found no evidence that the examined Gestalt principles can explain (un)crowding. Our results showed some advantages for symmetry, but this result should be interpreted with caution, especially considering that configurations that combined symmetry with another principle did not necessarily show such an advantage, as we discuss below. The rationale of our experiments is that when the central square is part of a group according to Gestalt principles, it should ungroup from the vernier and, hence, performance should be good. However, for each category of configurations, we found that some configurations showed better performance compared to the one-central-square condition, indicating uncrowding, while other conditions showed clear crowding, often even stronger than in the one-square condition. Performances within one category correlated as strongly as across categories (Figure 4 and Supplementary Figure 3). Performance for configurations with more than two Gestalt principles was, overall, not better than for those with one principle (e.g., configurations with CloseSymm mostly lead to strong crowding, see Figures 3, 4, CloseSymm). Often, the combination of principles (e.g., Figure 4, ContStret and CloseSymm) rather led to an increase in crowding than a release, contrary to the spirit of previous findings that showed better grouping when two principles are combined (Hochberg and Hardy, 1960; Ben-Av and Sagi, 1995; Kubovy and Wagemans, 1995; Quinlan and Wilton, 1998; Claessens and Wagemans, 2005, 2008; Kubovy and van den Berg, 2008; Oyama and Miyano, 2008; Luna and Montoro, 2011; Luna et al., 2016; Rashal et al., 2017a,b; Rashal and Kimchi, 2022). Still, crowding level (Thresh) and subjective grouping ratings (GlobRank, VStandRate, GStandRate, and GGroupRate) correlated significantly (Figure 5). Correlations were highest with the Vernier standout (VStand) ratings, which supports our claim that uncrowding happens when the vernier stands out from the flankers.

Finally, model simulations showed that grouping among high-level features had a stronger correlation with crowding level (Thresh) than low-level features (Figure 6). We did not simulate the physiological approaches' model performances as exploring physiological models was out of the scope of the current work. Additionally, numerous publications have attempted to explain (un)crowding performance under physiological frameworks (Manassi et al., 2016; Doerig et al., 2020a,b; Bornet et al., 2021a,b; Choung et al., 2021). For instance, Manassi et al. (2016), using the Fourier model, evaluated similar flanker configurations that consist of squares and stars and failed to explain (un)crowding. However, we admit that our configurations might have affected physiological-based models, such as in Waugh et al. (1993) and Mussap and Levi (1997), which used Fourier analysis and showed that Vernier acuity changes depending on the orientations of masks.

What are the implications of our results? There are several aspects. First, the physiological approach may not be tenable in its current form but is correct in principle. Maybe we need to give up the feedforward aspects and allow recurrent, complex interactions. For example, Doerig et al. (2019) have shown that one-stage, feedforward models cannot explain uncrowding since target information is irretrievably lost during feedforward processing. This holds true for local pooling models (e.g., Parkes et al., 2001; Solomon et al., 2004; Pelli, 2008; Greenwood et al., 2009, 2017; Dakin et al., 2010; Rosenholtz et al., 2012), and models that can account for global configurations, such as a Fourier model (Waugh et al., 1993; Mussap and Levi, 1997; Manassi et al., 2016),

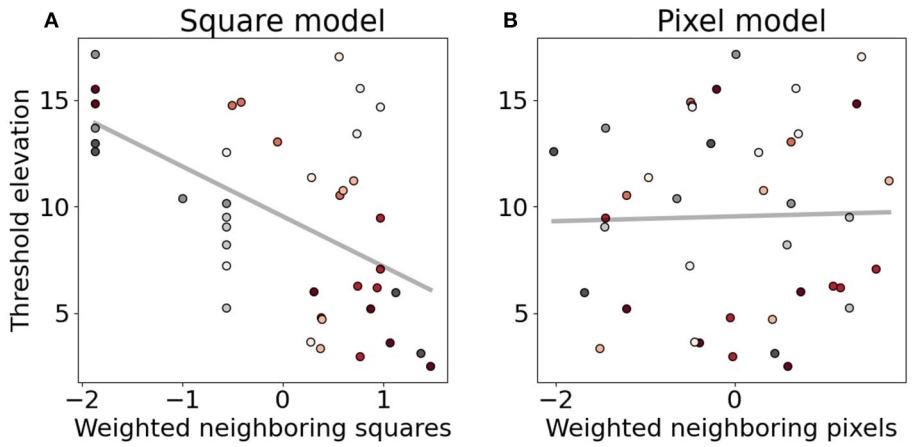


FIGURE 6
Correlations between model estimates and mean crowding level of each configuration for **(A)** the square model (higher-level feature) and **(B)** the pixel model (lower-level feature). The y-axis shows the mean threshold elevation, and the x-axis is the model estimates for each model. Dots represent configurations, the colors indicate Gestalt principles, corresponding to Figure 4.

epitomes model (for details see Jojic et al., 2003; Doerig et al., 2019), high dimensional feature pooling model (HD pooling; Rosenholtz et al., 2019; Bornet et al., 2021a; Choung et al., 2021), and deep networks (DNN; Doerig et al., 2020a). However, more complex models, employing recurrent processing, such as Capsule networks and the Laminart model, do not kill vernier-related information during feedforward and, therefore, can explain uncrowding results (Francis et al., 2017; Doerig et al., 2019, 2020b). In this models, indirect realism becomes even more indirect, including time-consuming, potentially idiosyncratic processing and the question arises whether these models adhere to spirit of the classic models.

Similar things can be said about the Gestalt approach. Indeed, the current Gestalt principles may be oversimplified. They work well for simple stimuli. However, future Gestalt cues may change the game (see for example, Todorović, 2011). For example, Gestalt principles may consider statistical principles, such as summary and ensemble statistics (Tiurina et al., 2022). In addition, it seems that our results do not argue against the Gestaltist's main credo: there is more in the mind than in the stimulus, and the whole is different from its parts. Perception is not one-to-one.

Maybe, these arguments are true. However, we think that the failures of both approaches show that there are deeper issues, related to the philosophical foundations of perception. As said, the Gestalt approach is rather silent about the external world because its main source of scientific reasoning comes from introspection, from how stimulus configurations look to us, and not from speculations about an external world, which is a latent variable because realism is indirect and, hence, we have no perceptual ground truth about it (Figure 1, upper panel). For this reason, Gestalt theory says very little about the world. Gestalt theory focuses on perception as a truly subjective science. However, detaching perception from an objective, mind-independent world opens up the possibilty that the Gestalt rules may be totally idiosyncratic. Hence, Gestalt theory loses its relationship to ground truth. This comes with quite some problems, in particular for an objective science of perception. From an evolutionary point of view, we may ask: why should there be more in the mind than what is in the world? Why should different people follow identical Gestalt principles if no constraints make some of the principles better than others?

Whereas the Gestalt approach is rather vague about its ontological commitments, the physiological approach clearly subscribes to indirect realism, includingthe ontological commitment to ordinary objects and a one-to-one mapping between the objects and their corresponding representations. As mentioned, mismatches are just unavoidable errors in the process because of the ill-posed problems of vision. However, the strong ontological commitment to the existence of everyday objects is not easily tenable. of course, one problem is that we can never verify this assumption since perception is indirect, i.e., we have no direct access to the world. However, we propose there is another main problem, namely, that the external world is much richer, i.e., there are much more fundamental entities (the physical particles), than mental representations. One can mathematically show that, in this situation, mind-independent ordinary objects cannot occur (Herzog and Doerig, 2021; Herzog, 2022). Apples are not the starting point of perception, they are the outcomes of perception. Perception is a mapping from fundamental physics directly to perception without an intermediate ontology of apples and alike. Hence, there is ground truth, as in the physiological approach, but the truth comes from physics (i.e., particles) as our primary source of knowledge, not by sensory or perceptual evidence of ordinary objects and a like. There is no accurate reconstruction of ordinary objects because there are no ordinary objects. In our view, the squares and disks are as mind-dependent as the cross in Figure 1.

In summary, as in the physiological approach, we propose that there is a mind-independent world of particles. Perception is a mapping from these particles into the world of mental representations, which are truly subjective in the spirit of the Gestaltists. For this reason, introspection is the tool of choice since there is no objectivity on the ordinary object level. Gestalt perception is realized by the neural wiring of each observer and hence may be fully idiosyncratic, i.e., different people do not

employ Gestalt principles in an identical manner. This is evident by manifest differences, as in the case of the #theDress. These differences are not unavoidable errors of a reconstruction process of ordinary objects but the unavoidable consequence of the truly subjective nature of vision, i.e., Gestalt vision. We are now ready for a happy marriage of both perspectives on perception.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Beritashvili Centre of Experimental Biomedicine, Georgia. The participants provided their written informed consent to participate in this study.

Author contributions

OHC, ER, and MHH designed the experiment, interpreted the analyzed data, and wrote the initial version of the manuscript. MK collected the data. OHC and MK analyzed data. OHC designed and built the mathematical model. All authors contributed to the manuscript revision, read, and approved the submitted version.

Funding

OHC, MK, and MHH were supported by the Swiss National Science Foundation (SNF) 320030_176153 Basics of visual processing: from elements to figures. ER was supported by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement no. 708007.

Acknowledgments

We thank Greg Francis for the valuable discussions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2023.1154957/full#supplementary-material

References

Bach, M. (1996). The freiburg visual acuity test-automatic measurement of visual acuity. *Optom. Vis. Sci.* 73, 49–53.

Bakdash, J. Z., and Marusich, L. R. (2017). Repeated measures correlation. *Front. Psychol.* 8, 456. doi: 10.3389/fpsyg.2017.00456

Barton, K. (2020). *MuMIn: Multi-Model Inference*. Available online at: https://CRAN.R-project.org/package=MuMIn (accessed September 1, 2023).

Bates, D., Machler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Software* 67, 1–48. doi: 10.18637/jss.v067.i01

Ben-Av, M. B., and Sagi, D. (1995). Perceptual grouping by similarity and proximity: experimental results can be predicted by intensity autocorrelations. *Vision Res.* 35, 853–866. doi: 10.1016/0042-6989(94)00173-J

Bornet, A., Choung, O-. H., Doerig, A., Whitney, D., Herzog, M. H., Manassi, M., et al. (2021a). Global and high-level effects in crowding cannot be predicted by either high-dimensional pooling or target cueing. *J. Vision* 21, 10–10. doi: 10.1167/jov.21.12.10

Bornet, A., Doerig, A., Herzog, M. H., Francis, G., and Van der Burg, E. (2021b). Shrinking Bouma's window: how to model crowding in dense displays. *PLOS Comput. Biol.* 17, e1009187. doi: 10.1371/journal.pcbi.1009187

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature* 226, 177–178. doi: 10.1038/226177a0

Bouma, H. (1973). Visual interference in the parafoveal recognition of initial and final letters of words. *Vision Res.* 13, 767–782. doi: 10.1016/0042-6989(73)90041-2

Bouma, H., and Andriessen, J. J. (1968). Perceived orientation of isolated line segments. *Vision Res.* 8, 493–507. doi: 10.1016/0042-6989(68)90091-6

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vis.* 10, 433–436. doi: 10.1163/156856897X00357

Chicherov, V., and Herzog, M. H. (2015). Targets but not flankers are suppressed in crowding as revealed by EEG frequency tagging. *NeuroImage* 119, 325–331. doi: 10.1016/j.neuroimage.2015.06.047

Chicherov, V., Plomp, G., and Herzog, M. H. (2014). Neural correlates of visual crowding. *NeuroImage* 93, 23–31. doi: 10.1016/j.neuroimage.2014.02.021

Choung, O-. H., Bornet, A., Doerig, A., and Herzog, M. H. (2021). Dissecting (un)crowding. *J Vision Res.* 21, 1–20. doi: 10.1167/jov.21.10.10

Chung, S. T. L., Levi, D. M., and Legge, G. E. (2001). Spatial-frequency and contrast properties of crowding. *Vision Res.* 41, 1833–1850. doi: 10.1016/S0042-6989(01)00071-2

Claessens, P. M. E., and Wagemans, J. (2005). Perceptual grouping in Gabor lattices: proximity and alignment. *Percep. Psychophysics* 67, 1446–1459. doi: 10.3758/BF03193649

Claessens, P. M. E., and Wagemans, J. (2008). A Bayesian framework for cue integration in multistable grouping: proximity, collinearity, and orientation priors in zigzag lattices. *Vision Res.* 8, 33–33. doi: 10.1167/8.7.33

Coates, D. R., Levi, D. M., Touch, P., and Sabesan, R. (2018). Foveal crowding resolved. *Sci. Rep.* 8, 1. doi: 10.1038/s41598-018-27480-4

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: L. Erlbaum Associates.

Dakin, S. C., Cass, J., Greenwood, J. A., and Bex, P. J. (2010). Probabilistic, positional averaging predicts object-level crowding effects with letter-like stimuli. *J. Vis.* 10, 14–14. doi: 10.1167/10.10.14

Danilova, M. V., and Bondarko, V. M. (2007). Foveal contour interactions and crowding effects at the resolution limit of the visual system. *J. Vis.* 7, 25–25. doi: 10.1167/7.2.25

Doerig, A., Bornet, A., Choung, O-. H., and Herzog, M. H. (2020a). Crowding reveals fundamental differences in local vs. Global processing in humans and machines. *Vision Res.* 167, 39–45. doi: 10.1016/j.visres.2019.12.006

Doerig, A., Bornet, A., Rosenholtz, R., Francis, G., Clarke, A. M., Herzog, M. H., et al. (2019). Beyond Bouma's window: how to explain global aspects of crowding? *PLOS Computat. Biol.* 15, e1006580. doi: 10.1371/journal.pcbi.1006580

Doerig, A., Schmittwilken, L., Sayim, B., Manassi, M., and Herzog, M. H. (2020b). Capsule networks as recurrent models of grouping and segmentation. *PLoS Comput. Biol.* 16, e1008017. doi: 10.1371/journal.pcbi.1008017

Flom, M. C., Heath, G. G., and Takahashi, E. (1963). Contour interaction and visual resolution: contralateral effects. *Science* 142, 979–980. doi: 10.1126/science.142.3594.979

Francis, G., Manassi, M., and Herzog, M. H. (2017). Neural dynamics of grouping and segmentation explain properties of visual crowding. *Psychol. Rev.* 124, 483–504. doi: 10.1037/rev0000070

Fründ, I., Haenel, N. V., and Wichmann, F. A. (2011). Inference for psychometric functions in the presence of non-stationary behavior. *J. Vis.* 11, 16–16. doi: 10.1167/11.6.16

Gadermann, A. M., Guhn, M., and Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: a conceptual, empirical, and practical guide. *Practical Asses. Res. Evaluat.* 17, 3. doi: 10.7275/n560-j767

Greenwood, J. A., Bex, P. J., and Dakin, S. C. (2009). Positional averaging explains crowding with letter-like stimuli. *Proc. Nat. Acad. Sci.* 106, 13130–13135. doi: 10.1073/pnas.0901352106

Greenwood, J. A., Szinte, M., Sayim, B., and Cavanagh, P. (2017). Variations in crowding, saccadic precision, and spatial localization reveal the shared topology of spatial vision. *Proceed. Nat. Acad. Sci.* 114, E3573–E3582. doi: 10.1073/pnas.1615504114

Han, S., Humphreys, G. W., and Chen, L. (1999). Parallel and competitive processes in hierarchical analysis: perceptual grouping and encoding of closure. *J. Exp. Psychol.* 25, 1411–1432. doi: 10.1037/0096-1523.25.5.1411

Herzog, M. H. (2022). The irreducibility of vision: gestalt, crowding and the fundamentals of vision. *Vision* 6, 35. doi: 10.3390/vision6020035

Herzog, M. H., and Doerig, A. (2021). *Why our Best Theories of Perception and Physics Undermine Realism*. Society for the Improvement of Psychological Science (SIPS) and the Center for Open Science (COS). doi: 10.31234/osf.io/r4sf9

Herzog, M. H., and Manassi, M. (2015). Uncorking the bottleneck of crowding: a fresh look at object recognition. *Curr. Opin. Behav. Sci.* 1, 86–93. doi: 10.1016/j.cobeha.2014.10.006

Herzog, M. H., Sayim, B., Chicherov, V., and Manassi, M. (2015). Crowding, grouping, and object recognition: a matter of appearance. *J. Vis.* 15, 5–5. doi: 10.1167/15.6.5

Herzog, M. H., Thunell, E., and Ögmen, H. (2016). Putting low-level vision into global context: Why vision cannot be reduced to basic circuits. *Vision Res.* 126, 9–18. doi: 10.1016/j.visres.2015.09.009

Hochberg, J., and Hardy, D. (1960). Brightness and proximity factors in grouping. *Percept. Mot. Skills* 10, 22. doi: 10.2466/PMS.10.1.22-22

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biomet. J.* 50, 346–363. doi: 10.1002/bimj.200810425

Jastrzębowska, M. A., Chicherov, V., Draganski, B., and Herzog, M. H. (2021). Unraveling brain interactions in vision: the example of crowding. *NeuroImage* 240, 118390. doi: 10.1016/j.neuroimage.2021.118390

Johnson, P. C. (2014). Extension of Nakagawa and Schielzeth's R2GLMM to random slopes models. *Methods Ecol. Evolut.* 5, 944–946. doi: 10.1111/2041-210X.12225

Jojic, N., Frey, B. J., and Kannan, A. (2003). "Epitomic analysis of appearance and shape," in *Proceedings Ninth IEEE International Conference on Computer Vision* (Nice: IEEE Computer Society), 34–34.

Kanizsa, G. (1979). *Organization in Vision: Essays on Gestalt Perception*. New York, NY: Praeger Publishers.

Kleiner, M., Brainard, D., and Pelli, D. (2007). *What's New in Psychtoolbox-3?* London: Pion Ltd.

Koffka, K. (1935). Principles of gestalt psychology, international library of psychology. *Philosophy Sci. Method* 32, 8.

Köhler, W. (1920). *Die physischen Gestalten in Ruhe und im stationaren Eine natur-philosophische Untersuchung [The physical Gestalten at rest and in steady state]*. Wiesbaden: Vieweg+Teubner Verlag. doi: 10.1007/978-3-663-02204-6_6

Kubovy, M., and van den Berg, M. (2008). The whole is equal to the sum of its parts: a probabilistic model of grouping by proximity and similarity in regular patterns. *Psychol. Rev.* 115, 131–154. doi: 10.1037/0033-295X.115.1.131

Kubovy, M., and Wagemans, J. (1995). Grouping by proximity and multistability in dot lattices: a quantitative Gestalt theory. *Psychol. Sci.* 6, 225–234. doi: 10.1111/j.1467-9280.1995.tb00597.x

Lev, M., Yehezkel, O., and Polat, U. (2014). Uncovering foveal crowding? *Scientific Rep.* 4, 1. doi: 10.1038/srep04067

Levi, D. M. (2008). Crowding—An essential bottleneck for object recognition: a mini-review. *Vision Res.* 48(5), 635–654. 009. doi: 10.1016/j.visres.2007.12.009

Lezama, J., Randall, G., Morel, J-. M., and Grompone von Gioi, R. (2016). Good continuation in dot patterns: a quantitative approach based on local symmetry and non-accidentalness. *Vision Res.* 126, 183–191. doi: 10.1016/j.visres.2015.09.004

Luna, D., and Montoro, P. R. (2011). Interactions between intrinsic principles of similarity and proximity and extrinsic principle of common region in visual perception. *Perception* 40, 1467–1477. doi: 10.1068/p7086

Luna, D., Villalba-García, C., Montoro, P. R., and Hinojosa, J. A. (2016). Dominance dynamics of competition between intrinsic and extrinsic grouping cues. *Acta Psychol.* 170, 146–154. doi: 10.1016/j.actpsy.2016.07.001

Malania, M., Herzog, M. H., and Westheimer, G. (2007). Grouping of contextual elements that affect vernier thresholds. *J. Vis.* 7, 1–1. doi: 10.1167/7.2.1

Manassi, M., Hermens, F., Francis, G., and Herzog, M. H. (2015). Release of crowding by pattern completion. *J. Vis.* 15, 16–16. doi: 10.1167/15.8.16

Manassi, M., Lonchampt, S., Clarke, A., and Herzog, M. H. (2016). What crowding can tell us about object representations. *J. Vis.* 16, 35. doi: 10.1167/16.3.35

Manassi, M., Sayim, B., and Herzog, M. H. (2012). Grouping, pooling, and when bigger is better in visual crowding. *J. Vis.* 12, 13–13. doi: 10.1167/12.10.13

Manassi, M., Sayim, B., and Herzog, M. H. (2013). When crowding of crowding leads to uncrowding. *J. Vis.* 13, 10–10. doi: 10.1167/13.13.10

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica* 22, 276–282. doi: 10.11613/BM.2012.031

Metzger, W. (1936). *Gesetze des Sehens [Laws of seeing]*. (Frankfurt am Main, Germany).

Metzger, W., Spillmann, L. T., Lehar, S. T., Stromeyer, M. T., and Wertheimer, M. T. (2006). *Laws of seeing*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/4148.001.0001

Mohr, D. L., and Marcon, R. A. (2005). Testing for a 'within-subjects' association in repeated measures data. *J. Nonparametric Stat.* 17, 347–363. doi: 10.1080/10485250500038694

Mussap, A. J., and Levi, D. M. (1997). Vernier acuity with plaid masks: the role of oriented filters in vernier acuity. *Vision Res.* 37, 1325–1340. doi: 10.1016/S0042-6989(96)00192-7

Nakagawa, S., Johnson, P. C. D., and Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J. Royal Soc. Interface* 14, 20170213. doi: 10.1098/rsif.2017.0213

Nakagawa, S., and Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods Ecol. Evolut.* 4, 133–142. doi: 10.1111/j.2041-210x.2012.00261.x

Oberfeld, D., and Stahn, P. (2012). Sequential grouping modulates the effect of non-simultaneous masking on auditory intensity resolution. *PLoS ONE* 7, e48054. doi: 10.1371/journal.pone.0048054

Overvliet, K. E., and Sayim, B. (2016). Perceptual grouping determines haptic contextual modulation. *Vision Res.* 126, 52–58. doi: 10.1016/j.visres.2015.04.016

Oyama, T., and Miyano, H. (2008). Quantification of Gestalt laws and proposal of a perceptual state-space model. *Gestalt Theory* 30, 29.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., and Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neurosci.* 4, 739–744. doi: 10.1038/89532

Pelli, D. G. (2008). Crowding: a cortical constraint on object recognition. *Curr. Opin. Neurobiol.* 18, 445–451. doi: 10.1016/j.conb.2008.09.008

Pelli, D. G., and Vision, S. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* 10.437–442. doi: 10.1163/156856897X00366

Quinlan, P. T., and Wilton, R. N. (1998). Grouping by proximity or similarity? Competition between the Gestalt principles in vision. *Perception* 27, 417–430. doi: 10.1068/p270417

R Core Team (2019). *R: A Language and Environment for Statistical Computing* (Vienna, Austria). R Foundation for Statistical Computing. Available online at: https://www.R-project.org/ (accessed September 1, 2023).

Rashal, E., and Kimchi, R. (2022). The competition between grouping cues can be resolved under inattention. *Attent. Percept. Psychophysics* 3, 1–12. doi: 10.3758/s13414-022-02576-2

Rashal, E., Yeshurun, Y., and Kimchi, R. (2017a). Attentional requirements in perceptual grouping depend on the processes involved in the organization. *Attent. Percept. Psychophysics* 79, 2073–2087. doi: 10.3758/s13414-017-1365-y

Rashal, E., Yeshurun, Y., and Kimchi, R. (2017b). The time course of the competition between grouping organizations. *J. Exp. Psychol. Human Percep. Perform.* 43, 608. doi: 10.1037/xhp0000334

Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research* (Evanston, Illinois). Northwestern University.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., and Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *J. Vis.* 12, 14–14. doi: 10.1167/12.4.14

Rosenholtz, R., Yu, D., and Keshvari, S. (2019). Challenges to pooling models of crowding: implications for visual mechanisms. *J. Vis.* 19, 15–15. doi: 10.1167/jov.19.7.15

Saarela, T. P., and Herzog, M. H. (2008). Time-course and surround modulation of contrast masking in human vision. *J. Vis.* 8, 23–23. doi: 10.1167/8.3.23

Saarela, T. P., and Herzog, M. H. (2009). Size tuning and contextual modulation of backward contrast masking. *J. Vis.* 9, 21–21. doi: 10.1167/9.11.21

Saarela, T. P., Sayim, B., Westheimer, G., and Herzog, M. H. (2009). Global stimulus configuration modulates crowding. *J. Vis.* 9, 5–5. doi: 10.1167/9.2.5

Saarela, T. P., Westheimer, G., and Herzog, M. H. (2010). The effect of spacing regularity on visual crowding. *J. Vis.* 10.17–17. doi: 10.1167/10.10.17

Sasaki, Y., Vanduffel, W., Knutsen, T., Tyler, C., and Tootell, R. (2005). Symmetry activates extrastriate visual cortex in human and non-human primates. *Proceed. National Acad. Sci. USA* 102, 3159–3163. doi: 10.1073/pnas.0500319102

Sayim, B., Manassi, M., and Herzog, M. (2014). How color, regularity, and good Gestalt determine backward masking. *J. Vision* 14, 8. doi: 10.1167/14.7.8

Sayim, B., Westheimer, G., and Herzog, M. H. (2008). Contrast polarity, chromaticity, and stereoscopic depth modulate contextual interactions in vernier acuity. *J. Vis.* 8, 12–12. doi: 10.1167/8.8.12

Sayim, B., Westheimer, G., and Herzog, M. H. (2010). Gestalt factors modulate basic spatial vision. *Psychol. Sci.* 21, 641–644. doi: 10.1177/0956797610368811

Sayim, B., Westheimer, G., and Herzog, M. H. (2011). Quantifying target conspicuity in contextual modulation by visual search. *J. Vis.* 11, 6–6. doi: 10.1167/11.1.6

Solomon, J. A., Felisberti, F. M., and Morgan, M. J. (2004). Crowding and the tilt illusion: toward a unified account. *J. Vis.* 4, 9–9. doi: 10.1167/4.6.9

Spearman, C. (1904). The proof and measurement of association between two things. *Am. JJ. Psychol.* 15, 72–101. doi: 10.2307/1412159

Strasburger, H., Harvey, L. O., and Rentschler, I. (1991). Contrast thresholds for identification of numeric characters in direct and eccentric view. *Percep. Psychophysics* 49, 495–508. doi: 10.3758/BF03212183

Taylor, M. M., and Creelman, C. D. (1967). PEST: efficient estimates on probability functions. *J. Acoust. Soc. America* 41, 782–787. doi: 10.1121/1.1910407

Tiurina, N. A., Markov, Y. A., Choung, O-.H., Herzog, M. H., and Pascucci, D. (2022). Unlocking crowding by ensemble statistics. *Curr. Biol.* 32, 4975–4981.e3. doi: 10.1016/j.cub.2022.10.003

Todorović, D. (2007). *W. Metzger, Laws of Seeing. Gestalt Theory* (Vienna: Krammer), 29, 176.

Todorović, D. (2011). What is the Origin of the Gestalt Principles? *Humana Mente*, 4, 17.

Toet, A., and Levi, D. M. (1992). The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Res.* 32, 1349–1357. doi: 10.1016/0042-6989(92)90227-A

Treder, M. S., and van der Helm, P. A. (2007). Symmetry vs. repetition in cyclopean vision: a microgenetic analysis. *Vision Res. 47*, 2956–2967. doi: 10.1016/j.visres.2007.07.018

van der Helm, P. A. (2014). "Symmetry perception," in *The Oxford Handbook of Perceptual Organization* ed J. Wagemans. Oxford University Press. doi: 10.1093/oxfordhb/9780199686858.013.056

von Ehrenfels, C. (1890). Über Gestaltqualitäten_About gestalt qualities. *Vierteljahrsschrift Für Wissenschaftliche Philosophie*, 14, 249–292.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., et al. (2012a). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychol. Bull.* 138, 1172. doi: 10.1037/a0029333

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., Van der Helm, P. A., et al. (2012b). A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychol. Bull.* 138, 1218. doi: 10.1037/a0029334

Waugh, S. J., Levi, D. M., and Carney, T. (1993). Orientation, masking, and vernier acuity for line targets. *Vision Res.* 33, 1619–1638. doi: 10.1016/0042-6989(93)90028-U

Wertheimer, M. (1912). Experimentelle studien uber das sehen von bewegung. *Z. Psychol.* 61, 481.

Wertheimer, M. (1922). Untersuchungen zur Lehre von der Gestalt I: Prinzipielle Bemerkungen [Investigations in Gestalt theory: I. *The general theoretical situation]. Psychol. Forsch.* 1, 47–58. doi: 10.1007/BF00410385

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt II. [Investigations in Gestalt Theory: II. Laws of organization in perceptual forms]. *Psychol. Forsch.* 4, 301–350. doi: 10.1007/BF00410640

Weymouth, F. W. (1958). Visual sensory units and the minimal angle of resolution. *Am. J. Ophthalmol.* 46, 102–113. doi: 10.1016/0002-9394(58)90042-4

World Medical Association (2013). World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA* 310, 2191–2194.

Zumbo, B. D., Gadermann, A. M., and Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for likert rating scales. *J. Mod. Appl. Stat. Methods* 6, 21–29. doi: 10.22237/jmasm/1177992180

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 21 September 2023
DOI 10.3389/fpsyg.2023.1243405

OPEN ACCESS

EDITED BY
Haluk Ogmen,
University of Denver, United States

REVIEWED BY
Birgitta Dresp-Langley,
Centre National de la Recherche Scientifique (CNRS), France
Matthew Self,
Netherlands Institute for Neuroscience (KNAW), Netherlands

*CORRESPONDENCE
Mary A. Peterson
✉ mapeters@arizona.edu

RECEIVED 28 June 2023
ACCEPTED 17 August 2023
PUBLISHED 21 September 2023

CITATION
Peterson MA and Campbell ES (2023) Backward masking implicates cortico-cortical recurrent processes in convex figure context effects and cortico-thalamic recurrent processes in resolving figure-ground ambiguity. *Front. Psychol.* 14:1243405. doi: 10.3389/fpsyg.2023.1243405

Backward masking implicates cortico-cortical recurrent processes in convex figure context effects and cortico-thalamic recurrent processes in resolving figure-ground ambiguity

Mary A. Peterson[1,2]* and Elizabeth Salvagio Campbell[1,2,3]

[1]Department of Psychology, University of Arizona, Tucson, AZ, United States, [2]Cognitive Science Program, University of Arizona, Tucson, AZ, United States, [3]College of Medicine Tucson, University of Arizona, Tucson, AZ, United States

Introduction: Previous experiments purportedly showed that image-based factors like convexity were sufficient for figure assignment. Recently, however, we found that the probability of perceiving a figure on the convex side of a central border was only slightly higher than chance for two-region displays and increased with the number of display regions; this increase was observed only when the concave regions were homogeneously colored. These convex figure context effects (CEs) revealed that figure assignment in these classic displays entails more than a response to local convexity. A Bayesian observer replicated the convex figure CEs using both a convexity object prior and a new, homogeneous background prior and made the novel prediction that the classic displays in which both the convex and concave regions were homogeneous were ambiguous during perceptual organization.

Methods: Here, we report three experiments investigating the proposed ambiguity and examining how the convex figure CEs unfold over time with an emphasis on whether they entail recurrent processing. Displays were shown for 100 ms followed by pattern masks after ISIs of 0, 50, or 100 ms. The masking conditions were designed to add noise to recurrent processing and therefore to delay the outcome of processes in which they play a role. In Exp. 1, participants viewed two- and eight-region displays with homogeneous convex regions (homo-convex displays; the putatively ambiguous displays). In Exp. 2, participants viewed putatively unambiguous hetero-convex displays. In Exp. 3, displays and masks were presented to different eyes, thereby delaying mask interference in the thalamus for up to 100 ms.

Results and discussion: The results of Exps. 1 and 2 are consistent with the interpretation that recurrent processing is involved in generating the convex figure CEs and resolving the ambiguity of homo-convex displays. The results of Exp. 3 suggested that corticofugal recurrent processing is involved in resolving the ambiguity of homo-convex displays and that cortico-cortical recurrent processes play a role in generating convex figure CEs and these two types of recurrent processes operate in parallel. Our results add to evidence that perceptual organization evolves dynamically and reveal that stimuli that seem unambiguous can be ambiguous during perceptual organization.

KEYWORDS

recurrent processing, figure-ground perception, context effects, ambiguity, thalamus, corticothalamic, cortico-cortical

Introduction

A central function of perception is segregating the visual field into foreground objects and their local backgrounds, yet the underlying mechanisms are not fully understood. Foreground-background perception (i.e., figure-ground perception) was long thought to result from low-level processes in a feedforward perceptual system. Evidence for this view was provided by demonstrations that figure assignment was determined by image-based cues such as convexity. For example, for stimuli like the one on the right in Figure 1A, a large majority of perceivers reported that the convex regions were the figures (e.g., Rubin, 1958; Hochberg, 1971; Pomerantz and Kubovy, 1986). Indeed, convexity was considered the principal figural prior (or "cue," Kanizsa and Gerbino, 1976). However, using displays exposed for 100 ms, Peterson and Salvagio (2008) found that the probability of perceiving convex regions as figures was only slightly above chance for two-region displays (like Figure 1A, left) and increased systematically with region number to 85–90% for eight-region displays like Figure 1A, right (see Figure 1E). These results indicated that the probability of perceiving convex regions as figures was boosted by global context, a factor that was not previously thought to have an influence. These global context effects (CEs) were observed only when the concave regions were homogeneous (as in Figures 1A, B), but not when they were heterogeneous (as in Figures 1C, D; see Figure 1E).

What processes produce these global CEs? The lack of convex figure CEs for Figures 1C, D ruled out grouping and probability summation, respectively (Peterson and Salvagio, 2008). Goldreich and Peterson (2012) replicated the global convex figure CEs with a Bayesian observer that incorporated a new background prior in addition to the convexity prior. They noted that backgrounds tend (more than figures) to be homogeneously colored. Consistent with this background prior, laboratory research shows that disconnected regions are more likely to be perceived as portions of a single surface when they are homogeneously rather than heterogeneously colored (Yin et al., 1997, 2000). The Bayesian observer also made the novel prediction that the classic eight-region displays like the one on the right in Figure 1A, in which both the convex and concave regions are homogeneously colored, are ambiguous during perceptual organization; ambiguity arises because the background prior of homogeneous color and the object prior of convexity oppose each other for convex regions. This prediction was surprising because the displays do not seem to be ambiguous: a large majority of observers report perceiving convex regions as figure (e.g., Kanizsa and Gerbino, 1976). If this prediction is confirmed, however, that will provide evidence that complex perceptual organization processes take place outside of awareness, even when a single prior was previously considered sufficient. Here, we report three experiments using backward pattern masks to examine the development of convex figure CEs for putatively ambiguous and unambiguous displays like those in Figures 1A, B, respectively, in order to better understand the dynamics of figure-ground segregation.

We are particularly interested in whether feedback from higher to lower levels in the visual hierarchy (i.e., recurrent processes) plays a role in convex figure CEs and in resolving ambiguity during perceptual organization. It is reasonable to assume that the homogeneous background prior entails perceptual completion, which seems to require feedback (Wyatte et al., 2012, 2014; Tang et al., 2014, 2018; for review see Thielen et al., 2019; Kreiman and Serre, 2020). It is known that contextual influences on neural responses are mediated by recurrent processing (e.g., Lamme, 1995; Zipser et al., 1996; Gilbert and Li, 2013). Recurrent processes within the primate cortex modulate the responses of V1 neurons to figures defined by contrasting features inside and outside their receptive fields (e.g., Lamme and Roelfsema, 2000; Lamme et al., 2002; Craft et al., 2007; see Kelly and Grossberg, 2000; Jehee et al., 2007 for models implementing recurrent processes in figure-ground perception). Recently, Self et al. (2019) showed that recurrent input to V1 from a higher cortical level plays a role in resolving a local ambiguity in figure-ground organization. Going beyond cortico-cortical recurrent processing, Sillito and Jones (2002), Jones et al. (2015), and Poltoratski et al. (2019) found that cortico-fugal feedback modulates the neural representation of figures in the primate thalamus. Indeed, cortico-thalamic feedback seems to be automatic; Jones et al. (2015) hypothesized that it iteratively refines local thalamic responses to be consistent with global responses in higher-level cortical areas. Based on this previous research regarding context effects in figure-ground perception, we investigated whether recurrent processing plays a role in convex figure CEs.

We began by investigating the development of convex figure CEs obtained with the classic displays like those in Figure 1A; see also Figures 2A–C. Convex figure CEs are characterized by substantially higher convex figure reports for eight-region than two-region displays. In the eight-region displays used in all experiments in this article, the concave regions were homogeneous, an essential ingredient for convex figure CEs. In the classic displays used in Exp. 1, the convex regions were also homogeneous. In the displays tested In Exp. 2, the convex regions were heterogeneous. Henceforth, these two types of displays will be labeled *homo-convex* and *hetero-convex* displays, respectively. Test displays were exposed for 100 ms (the duration used by Peterson and Salvagio, 2008) and were followed by a 200-ms pattern mask after interstimulus intervals (ISIs) of 0, 50, or 100 ms. The 100-ms duration during which the test displays were shown is sufficient for feedforward activation through the visual hierarchy (e.g., Lamme and Roelfsema, 2000; Bullier, 2001). Hence, activation from the mask is unlikely to interfere with feedforward activation from the display (e.g., Lamme et al., 2002; Breitmeyer and Ogmen, 2006; Roelfsema, 2006; Di Lollo, 2007; Fahrenfort et al., 2007; Wyatte et al., 2012, 2014; but see Breitmeyer and Ogmen, 2022). However, feedforward activation from a subsequently presented pattern mask can add noise to the substrate for recurrent processing initiated by a preceding stimulus. Perceptual organization that depends on recurrent processes would emerge more slowly as a consequence. Therefore, if recurrent processing is involved in convex figure CEs, the probability of observing CEs should increase with display-to-mask ISI. Moreover, if, as hypothesized, *homo-convex* displays are ambiguous and ambiguity resolution also requires recurrent processes, convex figure CEs may emerge in a longer ISI condition for *homo-* than *hetero-convex* displays. This is because it takes time to resolve ambiguity (Peterson and Lampignano, 2003; Peterson and Enns, 2005; Brooks and Palmer, 2011). The outcome of these experiments will yield insights into the complex interactive

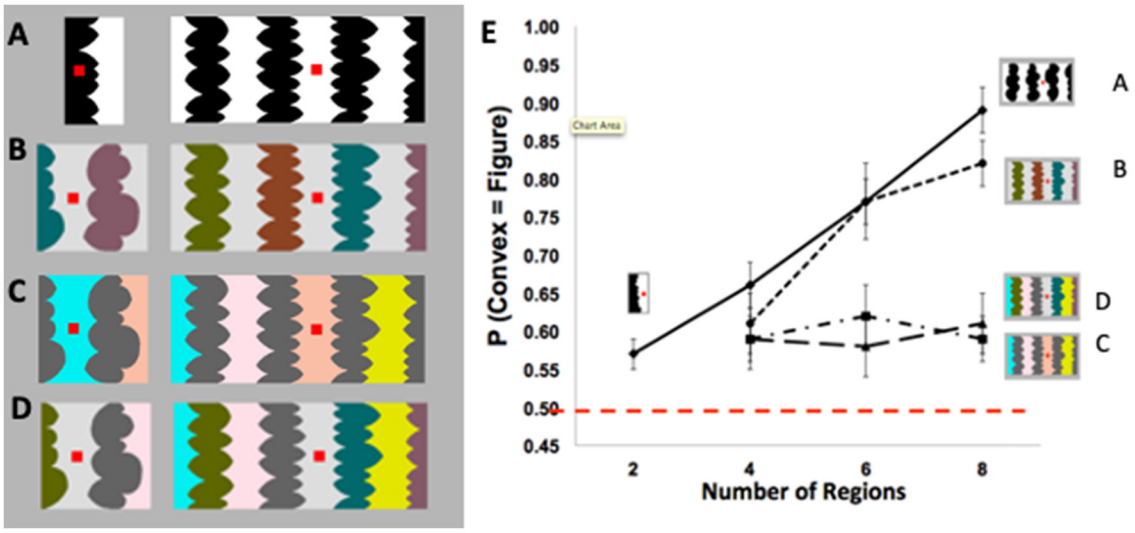


FIGURE 1
(A) Two- (left) and eight-region (right) displays with homogeneous (homo) convex and homo-concave regions. **(B–D)** Four- (left) and eight-region (right) displays comprising **(B)** heterogeneous (hetero) convex and homo-concave regions; **(C)** homo-convex and hetero-concave regions; **(D)** hetero-convex and hetero- concave regions. **(E)** Proportion of convex figure reports as a function of region number for unmasked 100-ms exposures of displays **(A–D)**. P(convex = figure) reports increased with region number only when concave regions were homogeneous (these figures are adapted from Figures 2–5 in Peterson and Salvagio, 2008). Participants' task was to report whether the red probe appeared "on" or "off "the region they perceived as the figure at the nearest border. The dashed red line indicates chance performance (50% convex figure reports). Error bars represent standard error of the mean.

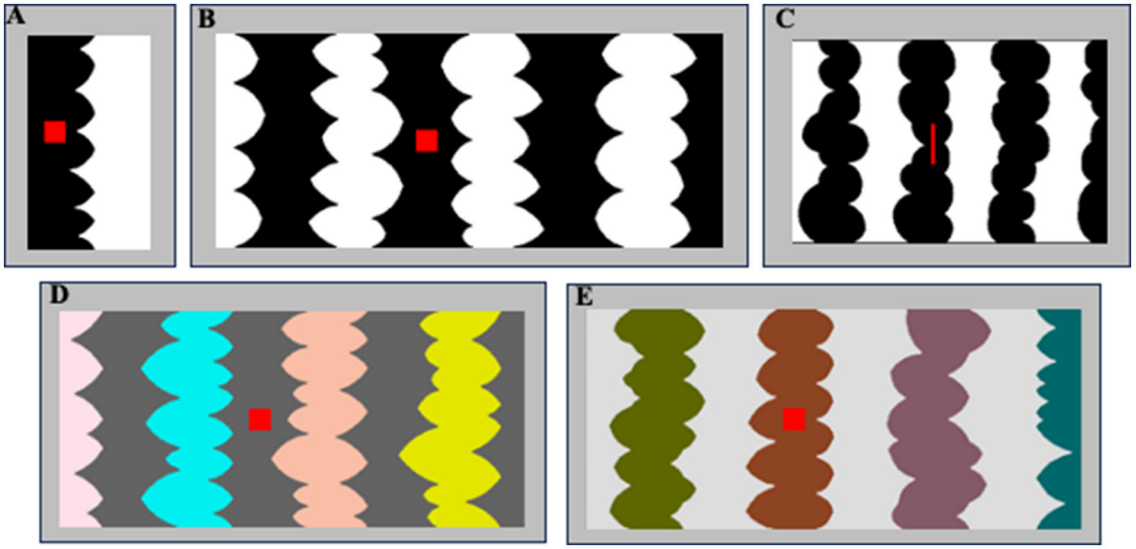


FIGURE 2
(A, B) Sample two- and eight-region homo-convex displays used in Exp. 1A. Convex region(s) are black in **(A)** and white in **(B)**; located to the left of the central border in **(A)** and to the right of the central border in **(B)**. The red probe is on the convex region in **(A)** and off the convex region in **(B)**. **(C)** A sample eight-region display used in Exp. 1B with black convex regions. The red probe is on the convex region to the left of the central border. **(D, E)** Sample eight-region hetero-convex displays used in Exp. 2. Convex regions are HL in **(D)** and LL in **(E)**.

processes that lead to the determination of where convex objects lie with respect to borders in scenes.

Experiment 1

In Exp. 1A, participants viewed two- or eight-region *homo-convex* displays like those in Figures 1A, 2A, B for 100 ms; for each display they reported whether they perceived the convex region as a figure. Displays were followed by a pattern mask at one of three ISIs (0, 50, or 100 ms). Convex figure CEs are defined by significantly higher convex figure reports for eight-region than two-region displays. We found that convex CEs increased in magnitude as the display-to-mask ISI increased from 0 to 100 ms, consistent with predictions if recurrent processing is involved in generating convex figure CEs. In Exp. 1B, we presented narrower eight-region

displays (see Figure 2C) in the same display-to-mask ISI conditions used in Exp 1A to investigate whether the recurrent processes implicated in convex figure CEs operate between levels of the visual hierarchy (vertically) or within a level (i.e., horizontally). In both cases, backward pattern masks could add noise to the substrate for recurrent processes. However, within-level recurrent processes take more time as the distance they must travel increases whereas those between levels are substantially less affected by distance (Girard et al., 2001). Therefore, horizontal within-level recurrent processes would be implicated if convex figure CEs emerge at a shorter ISI for narrow displays than for wider displays, whereas vertical between-level recurrent processes would be implicated if convex figure CEs develop along the same time course for narrow and wide displays.

Participants

Participants in Exp. 1 and all experiments reported in this article were undergraduate students at the University of Arizona who took part to partially fulfill the requirements of an introductory Psychology class. They signed a consent form approved by the University of Arizona IRB before participating. All participants reported normal or corrected-to-normal vision. The data from participants who failed to respond within 3,000 ms on at least 85% of the trials were removed. This standard criterion was applied to all conditions in all experiments. These aspects of the experiments held true for all participants in all experiments reported in this article.

A total of 200 students took part in Exp. 1A; 104 students participated in the follow-up experiment; and 104 students (59 F; 37M) participated in Exp. 1B. The number of participants whose data were removed because they did not meet the standard criterion was eight for Exp. 1A, four for the follow-up experiments to Exp. 1A, and eight for Exp. 1B.

Stimuli

The stimuli used in Exp. 1A were 112 two- and eight-region *homo-convex* displays (56 per region number condition) comprising alternating low luminance (LL; RGB = 0,0,0) and high luminance (HL; RGB = 255,255,255) convex and concave regions (see Figures 1A, 2A–C). In Exp. 1A, the stimuli were all equal in height (5.65°H) and varied in width (W): Two-region displays were on average 2.92°W (range: 2.45–3.28°; see Figure 2A); eight-region displays were 13.87°W (range: 12.17–15.87°; see Figure 2B). In Exp. 1B, the stimuli were 96 eight-region *homo-convex* displays that were 5.53°H x 8.53°W (see Figure 2C). Regions were deemed convex if their parts, delimited by successive minima of curvature, had positive curvature (cf., Peterson and Salvagio, 2008). Convex regions were LL and concave regions were HL in half of the displays, with achromatic colors reversed in the remaining half. In half the displays, the region to the right of the central border was convex; in the other half, the region to the right of the central border was concave (see Peterson and Salvagio, 2008 for complete stimulus construction details). An invisible rectangular frame around the displays cut the leftmost and rightmost regions of the displays in half, giving the impression that they continued behind the frame. Burrola and Peterson (2014) and Mojica and Peterson (2014) found that without a frame that allows perceptual completion, CEs are not observed.

A red probe was centered vertically on the region to the right or left of the central border. The red probe was a square in Exp. 1A and a narrow bar in Ex. 1B because the individual regions of the narrow displays were necessarily narrower (see Figure 2C). In previous experiments, responses to square and bar probes did not differ (Peterson and Salvagio, 2008).

Displays were centered on a medium gray backdrop (RGB = 182, 182, 182; luminance = 11.95 ft-L) that filled the screen (17.7°H x 22.8°W) of a 21-in Sony CRT monitor. The HL and LL regions were equal luminance steps below and above the backdrop; hence, contrast with the backdrop did not serve as a depth cue (see O'Shea et al., 1994). The masks used in Exp. 1 comprised a geometric pattern with white, black, and medium gray regions. In Exp. 1A, the masks were 5.83°H and were 2.98°W for two-region displays and 16.15°W for eight-region displays. In Exp. 1B, the masks were 5.53°H x 9.69°W. A sample mask for *homo-convex* displays is shown in Figure 3.

Design and procedure

In all experiments in this article, conditions were tested between-subjects to avoid contamination of one condition by another. Participants were assigned via a Latin square to a single region number and ISI condition when they arrived at the laboratory. After signing the consent form, participants were instructed on the nature of figure-ground perception and their task using instructions displayed on the computer; an experimenter read these instructions aloud while they were displayed and stayed in the room during practice trials to answer any questions.

Each trial began with a fixation cross, centered where the central edge of the upcoming test display would be located. Participants were instructed to fixate their eyes on this cross and to press the foot pedal when they were ready to begin each trial. Upon pressing the foot pedal, a single display was presented for 100 ms. The pattern mask (200 ms) was presented 0, 50, or 100 ms after the experimental display. Figure 3 illustrates the trial sequence. The presentation software automatically advanced to the next trial when participants responded or after 3,000 ms had elapsed (a time-out was recorded if participants did not respond within the 3,000-ms window). Viewing distance was constrained by a chinrest mounted 96 cm from the monitor.

On each trial in Exp. 1, a *homo-convex* test display appeared for 100 ms. Participants' task was to report whether the red probe on the display was located "on" or "off" the region they perceived as the figure shaped by the nearest border. This probe on/off task provides a valid and reliable index of figure assignment near a border (e.g., Hoffman and Singh, 1997; Peterson and Salvagio, 2008; Mojica and Peterson, 2014; Peterson et al., 2017). The instructions stated that there were no correct answers in the experiment, that different people see the displays differently, and that the experimenters were interested in participants' first impression of

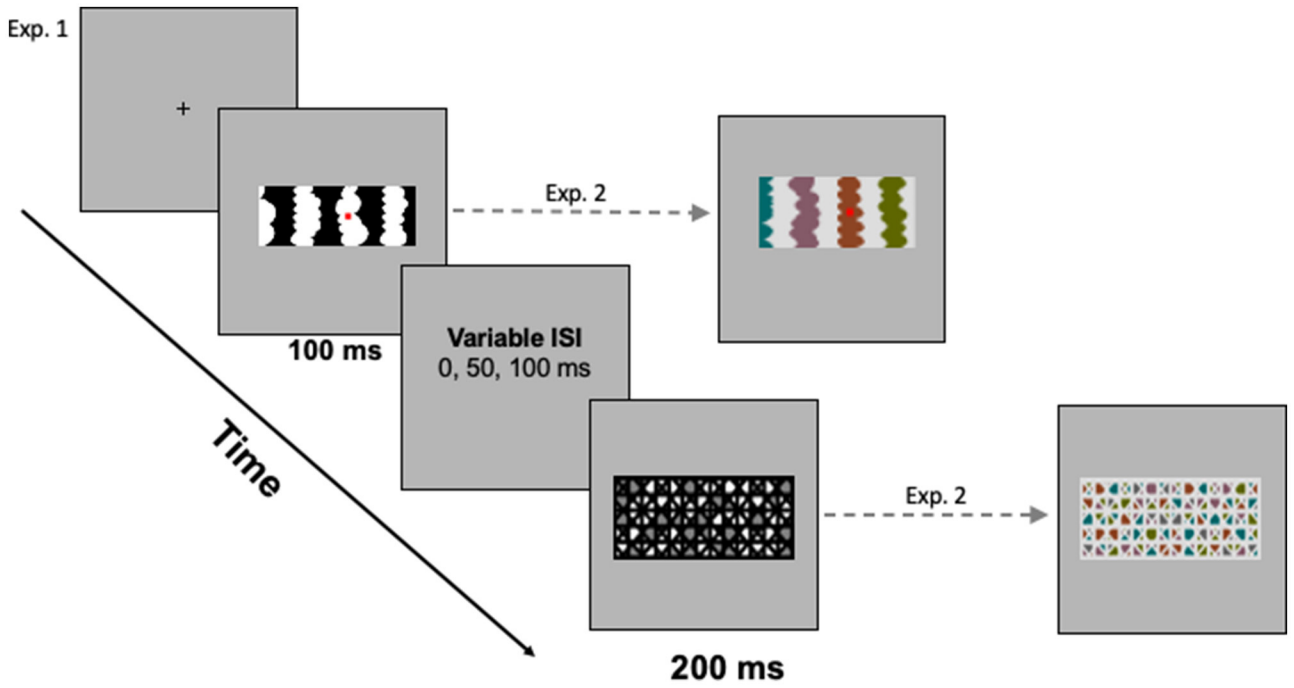


FIGURE 3
(Left) The trial sequence with a sample mask for *homo-convex* displays. **(Right)** Sample stimulus and mask for *hetero-convex* displays. A blank gray screen was shown for a variable interstimulus interval (ISI) between the test display and the mask.

the display. Participants were told that a random pattern would appear after the test display disappeared (this was the mask) and that they only had to look at it this pattern, not respond to it.

On experimental trials in Exp. 1A, each participant viewed 56 randomly presented trial unique *homo-convex* displays in one region number (two- or eight-region) and display-to-mask ISI condition. Participants in Exp. 1B viewed 96 trial-unique *homo-convex* displays. Participants made their on/off judgment regarding the red probe by pressing the top or bottom button on a custom button box. Assignment of buttons to "on"/"off" responses was balanced across subjects. Before the experimental trials, participants completed eight practice trials; none of the displays used in the practice trials appeared in the experimental trials. Participants were left alone to complete the experimental trials.

Data analysis

The proportion of trials on which the convex region closest to fixation was perceived as the figure/object was calculated for each participant by summing the number of trials on which they reported "on" when the probe appeared on the convex region, and "off" when the probe appeared on the concave region and dividing this sum by the total number of trials on which they responded (i.e., excluding timeouts and responses faster than 200 ms).

Results

Experiment 1A

As can be seen in the black and white bars in Figure 4, convex figure reports increased with region number, $F_{(1, 186)} = 19.39$, $p < 0.001$, $\eta^2 = 0.094$ and with display-to-mask ISI, $F_{(2, 186)} = 4.98$, $p < 0.009$; $\eta^2 = 0.051$. Importantly, an interaction between region number and ISI, $F_{(2, 186)} = 3.06$, $p < 0.05$; $\eta^2 = 0.032$, showed that convex figure reports increased with ISI for eight-region displays, $F_{(2, 93)} = 4.894$, $p = 0.01$, $\eta^2 = 0.095$, but not for two-region displays, $F < 1$. To represent the magnitude of the convex figure CEs, the difference between convex figure reports for eight- vs. two-region displays was calculated for each ISI condition. This *CE index* was not statistically different from zero in the 0-ms ISI condition [0.033, $F_{(1, 62)} = 1.629$, $p > 0.20$]; it just reached statistical significance in the 50-ms ISI condition [0.086, $F_{(1, 62)} = 4.12$, $p < 0.05$] and was robust in the 100-ms ISI condition [0.16, $F_{(1, 62)} = 16.43$, $p < 0.001$]. The *CE index* was statistically higher in the 100-ms than the 50-ms display-to-mask ISI condition, $p < 0.002$ (see Table 1).

Follow-up experiment

To investigate whether convex figure CEs continue to develop longer than 100-ms after the offset of the test stimulus, we presented different groups of participants two- and eight-region displays in 200-ms and 300-ms display-to-mask ISI conditions. We compared convex figure responses in these new conditions to those reported in the 100-ms ISI condition of Exp. 1A in a 2 (region number) X

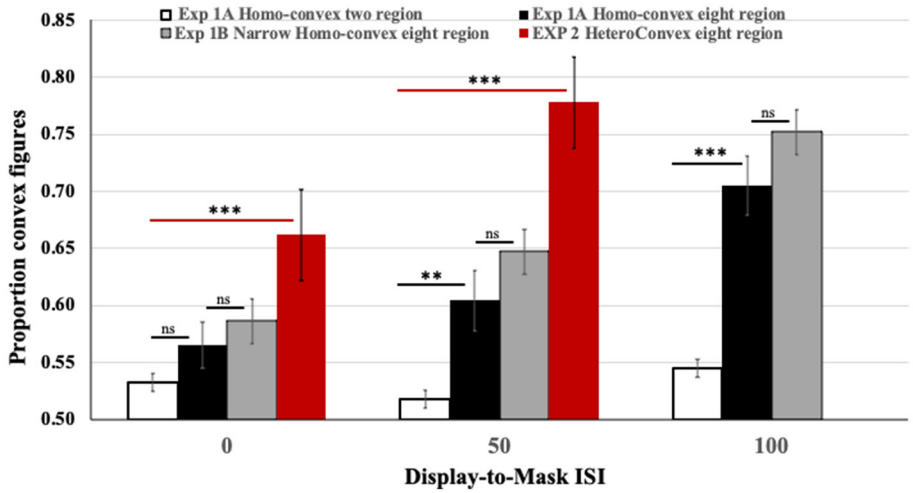


FIGURE 4
Results of Exps. 1A, 1B, and 2. Black and white: The proportion of convex figure reports for two-region (white) and eight-region (black) homo-convex displays in Exp. 1A. Gray: The proportion of convex figure reports for narrower eight-region homo-convex displays in Exp. 1B. Red: The proportion of convex figure reports for eight-region hetero-convex displays in Exp. 2. Black horizontal lines indicate differences between results for two- and eight-region displays in Exp. 1A. Gray horizontal lines indicate ns differences between results for eight region displays in Exps. 1A and 1B. Red horizontal lines indicate differences between results for two- and eight-region displays in Exp. 2. $^{**}p < 0.05$; $^{***}p < 0.001$; ns, no significant difference. Error bars represent standard errors.

TABLE 1 The proportion of convex figure reports and CE indices as a function of region number and display-to-mask ISI in Exp. 1A and the follow-up Experiment.

	Display-to-mask ISI (ms)				
	0	50	100	200	300
	Proportion convex figures				
8-region	0.57	0.60	0.71	0.64	0.71
2-region	0.53	0.52	0.55	0.49	0.53
	CE index				
	0.03	0.09*	0.16***	0.15***	0.18***

The CE Index is the difference between convex figure reports for eight- and two-region displays.
CE, Context Effect.
$^{*}p < 0.05$; $^{***}p < 0.01$.

3 (ISI) ANOVA. A significant main effect of region number was observed, $F_{(1, 159)} = 33.219$, $p < 0.001$, $\eta^2 = 0.0174$, but there was no effect of ISI, $F_{(2, 159)} = 2.002$, $p > 0.13$ (see Table 1).

Together with the follow-up experiments, the results of Exp 1 show that convex figure CEs reached asymptote for 100-ms displays in the 100-ms display-to-mask ISI condition—200 ms after stimulus onset. It is plausible that pattern masks shown 0 and 50 ms after the offset of a 100-ms stimulus (and maybe longer up to 100 ms) interfere with recurrent processing following the initial analysis of the test display, thereby preventing the emergence of convex figure CEs. We continue to investigate the feasibility of this interpretation in subsequent experiments reported in this article. At this point, an explanation holding that the test display-off signal is critical for convex figure CEs and that masks interfere with that signal in the 0-ms display-to-mask ISI condition remains possible (Macknik and Martinez-Conde, 2007); it is shown to be infeasible by the results of Exp. 2.

Experiment 1B

To better characterize recurrent processes implicated by the results of Exp. 1A, we compared convex figure reports obtained for narrow eight-region displays in the three display-to-mask ISI conditions to those obtained for the wider eight-region displays in Exp. 1A. The gray bars in Figure 4 show the results. A 2(display width) X 3(ISI) ANOVA showed a main effect of ISI: Convex figure reports increased as display-to-mask ISI increased, $F_{(2, 186)} = 13.354$, $p < 0.001$, $\eta^2 = 0.125$. Neither a main effect of display width, $F_{(1, 186)} = 2.197$, $p = 0.140$, $\eta^2 = 0.01$, nor an interaction between display width and display-to-mask ISI was observed, $F < 1.0$. The finding that convex figure CEs for narrow and wide displays showed the same developmental trajectory over variations in display-to-mask ISI characterizes the recurrent processes as operating between levels of the visual hierarchy rather than within a level (i.e., vertically rather than horizontally).

Experiment 2

In Exp. 1, we found statistically significant convex figure CEs for *homo-convex* displays in the 50-ms display-to-mask ISI condition and larger convex figure CEs in the 100-ms ISI condition. That masks shown up to 100 ms after stimulus offset interfered with the generation of convex figure CEs is consistent with the hypothesis that recurrent processes play a role. Recall, however, that it has been proposed that *homo-convex* displays are ambiguous because when convex regions are homogeneous the convexity object prior and the homogeneous background prior oppose each other. If this proposal is correct, Exp. 1 may have assessed the need for recurrent processing in ambiguity resolution as well as in convex figure CEs.

In Exp. 2, we used the same procedure with *hetero-convex* displays. *Hetero-convex* displays are unambiguous because

disconnected heterogeneous convex regions are unlikely to be completed into a single surface (especially given that they change color only when out of sight in *hetero-convex* displays; Yin et al., 1997, 2000; Goldreich and Peterson, 2012). Only the homogeneous concave regions of *hetero-convex* displays would support perceptual completion into a background. The colored interior and the borders of heterogeneously colored convex regions are likely to be combined when convex figures are perceived (cf., Grossberg and Mingolla, 1985; Kellman and Shipley, 1991; Zhou et al., 2000) but this combination is insufficient for convex figure CEs; homogeneously colored concave regions are necessary (see Figure 1). Exp. 2 will provide evidence regarding whether convex figure CEs *per se* grow with display-to-mask ISI. In addition, a finding that convex figure CEs emerge and reach asymptote in a shorter display-to-mask ISI condition in Exp. 2 for *hetero-* than *homo-convex* displays in Exp. 1 will support the hypothesis that *homo-convex* displays are ambiguous. If that result is obtained, the difference between the ISIs in which equivalent convex figure reports are obtained in the two conditions may estimate how much time is required to resolve the ambiguity of *homo-convex* displays.

Participants

A total of 67 participants (52F; 15M) were tested in Exp. 2. When they entered the laboratory, they were assigned via an ABBA order to one of two display-to-mask ISI conditions: 0 or 50 ms. The data from two participants were excluded from the analysis because they did not meet our response rate criterion. The data from one additional participant were excluded because they pressed the response button immediately after pressing the foot pedal. A total of 42 participants (28 F) took part in a follow-up experiment in which displays were exposed for 80 ms and were followed immediately by a 200-ms mask. They were assigned via an ABBA procedure to view either the same eight-region displays viewed by participants in Exp. 2 or 56 two-region displays used in Experiment 1A.

Stimuli

The stimuli used in Exp. 2 were 64 eight-region displays from Peterson and Salvagio (2008, Exp. 3): in half the displays *hetero-convex* regions alternated with *homo-concave* regions; these were the experimental stimuli. In the other half, *hetero-convex* regions alternated with *hetero-concave* regions; these were filler stimuli included to reduce tendencies to form a strategy of always reporting either the *homo* or the *hetero* regions as figures (cf. Peterson and Salvagio, 2008). The choice of which 32 of the 64 displays served as filler stimuli was balanced across participants. As per Peterson and Salvagio, responses to the filler stimuli were not analyzed.

The displays were all equal in height (5.65°) and varied in width, subtending a mean visual angle of 13.59° (range: 11.54–15.65°). The convex regions differed in color. The convex region sharing the central border with the concave region was always gray. The other convex regions were filled with one of four colors: yellow, magenta, cyan, or orange. These colors appeared once per display, and across displays appeared on each of the remaining three convex regions equally often. The concave regions were filled with either HL or LL gray. The convex and concave regions differed in contrast polarity: when the concave regions were HL, the convex regions were LL and vice versa. Samples are shown in Figures 2D, E (as in the other experiments, stimuli were shown on a medium gray backdrop).

In the filler displays, the alternating regions were HL or LL and colored gray, yellow, magenta, or cyan. The two central regions were always filled with HL and LL gray; hence, the central regions in the two types of displays were equated. The remaining colors were used to fill the other regions. The same color was never used to fill two consecutive regions; nor was it used in multiple convex (or concave) regions in a single display. Convex regions were HL in half the displays and LL in the rest. The convex and concave regions differed in contrast polarity: when the luminance of the concave regions was high, that of convex regions was low and vice versa. Michelson contrast at the central border = 0.72. Michelson contrasts at the other borders ranged from 0.62 to 0.78.

The mask that followed the figure-ground display consisted of a geometric pattern that measured 6.0° H x 17.7° W (samples are shown in Figures 2, 3). A mask composed of LL gray and HL colored regions followed displays where the concave regions were LL-gray and the convex regions were HL colors and a mask composed of HL gray and LL colored regions followed displays where the concave regions were HL-gray and the convex regions were LL colors. HL and LL masks followed the filler displays equally often.

Procedure

In Exp. 2, the trial structure was the same as in Experiment 1 (see Figure 3). Test displays were exposed for 100 ms and followed by a 200-ms mask after an ISI of 0 or 50 ms. The filler displays were randomly intermixed with the *hetero-convex* displays. In other respects, the apparatus and procedure of Exp. 2 were the same as that of Exp. 1. In the follow-up experiment, two- and eight-region displays were exposed for 80 ms and followed immediately by a 200-ms mask.

Results and discussion

The results obtained with eight-region *hetero-convex* displays in Exp. 2 are shown in red in Figure 4. To assess how convex figure CEs for *hetero-convex* displays were affected by display-to-mask ISI, convex figure reports obtained for eight-region *hetero-convex* displays in Exp. 2 were first compared to those obtained with two-region displays in Exp. 1 (with only one convex region, two-region displays cannot be classified as either *homo-* or *hetero-convex*). The ANOVA showed a main effect of region number, $F_{(1, 124)} = 45.838$, $P < 0.001$, $\eta^2 = 0.270$ and an interaction between region number and ISI, $F_{(1,124)} = 5.077$, $p = 0.026$, $\eta^2 = 0.039$. Convex figures reports for eight-region displays increased with display-to-mask ISI (as in Exp. 1), whereas convex figure reports for two-region displays did not. The results of Exp. 2 are consistent with the interpretation that convex figure CEs entail recurrent processes.

Unlike Exp. 1, in Exp. 2, convex figures were perceived significantly more often in eight-region than two-region displays in the 0-ms ISI condition, $F_{(1, 62)} = 10.738$, $p = 0.002$, $\eta^2 = 0.148$ (the CE index of 0.13 was significantly >0, $p < 0.001$). This finding is inconsistent with a claim that the absence of convex figure CEs for homo-convex displays in the 0-ms display-to-mask ISI condition of Exp. 1 can be explained by mask-induced interference with a display offset signal as per Macknik and Martinez-Conde (2007). It also suggests that the processes generating convex figure CEs for hetero-convex displays are underway while the test displays are exposed. Replicating Exp. 1, the difference between convex figure reports for eight- and two-region displays was larger in the 50-ms display-to-mask ISI condition (CE index of 0.25), indicating that convex figure CEs for *hetero-convex* displays continue to develop after display offset. Because masks shown 50 ms after a 100-ms display are highly unlikely to interfere with feedforward processing, these results are consistent with the hypothesis that recurrent processes play a role in generating convex figure CEs.

Follow-up to Exp. 2

The finding that convex figure CEs were evident in the 0-ms display to-mask ISI condition raised the question of when CEs fist emerge for *hetero-convex* displays. To address this question, we showed 80-ms two-region displays and eight-region *hetero-convex* displays to different groups of participants in a 0-ms display-to-mask ISI condition. No convex figure CEs were observed: convex figures were perceived on statistically equivalent proportions of trials for two- and eight-region displays: 0.56 and 0.60 [$F_{(1, 36)} < 1$; the CE index was 0.04]. Together, the results of Exp. 2 and this follow-up experiment suggest that, when object and background priors are not in opposition for convex regions, the processes that produce convex figure CEs in eight-region displays take more than 80 ms after display onset and continue for up to 150 ms (i.e., 50 ms after the offset of the 100-ms display). These findings accord with previous estimates of how long perceptual completion takes although our displays are different from those examined by previous authors (e.g., Sekuler and Palmer, 1992; Ringach and Shapley, 1996; Guttman et al., 2003).

Why do convex figure CEs emerge earlier in time for *hetero-convex* displays (Exp. 2) than for *homo-convex* displays (Exp. 1)? We have attributed this temporal difference to processes that resolve the ambiguity of *homo-convex* displays. This raises the question of whether ambiguity resolution occurs in parallel with the generation of alternative interpretations for *homo-convex* displays or whether it occurs in a later decision process. It is reasonable to assume that perceptual completion processes generating background interpretations for homogeneous regions are underway while *homo-convex* displays are exposed as well as while *hetero-convex* displays are exposed, yet convex figure CEs are evident in convex-figure responses 50 ms later for *homo-* than for *hetero-convex* displays. We found that the cost of ambiguity resolution is approximately constant with increases in the ISI between eight-region displays and the subsequent backward masks: AN ANOVA comparing convex figure reports for eight region displays in the 0- and 50-ms ISI conditions common to both Exps. 1A and 2 showed a main effect of display type, $F_{(1, 124)} = 15.681$, $p < 0.001$, $\eta^2 = 0.112$ (higher convex figure reports for hetero- than homo-convex displays); and a main effect of ISI, $F_{(1, 124)} = 5.11$, $P = 0.026$, $\eta^2 = 0.04$ (higher convex figure reports in the 50-ms ISI condition than the 0-ms ISI condition). No interaction between display type and ISI was observed, $F_{(1, 124)} = 1.273$. This analysis reveals that the disadvantage for convex figure reports in eight-region *homo-* vs. *hetero-convex* displays is present in the 0-ms condition and remains stable as convex figure CEs develop. This result is consistent with the interpretation that ambiguity resolution processes operate in parallel with the processes generating convex figure CEs. Indeed, evidence for convex figure CEs in *homo-convex* displays lagged behind evidence for convex figure CEs in *hetero-convex* displays by ~50 ms: Convex figure reports for eight-region homo-convex displays in the 50-ms ISI condition where convex figure CEs first emerged in Exp. 2 (mean: 0.60; se: 0.04) were statistically equivalent to convex figure reports for eight-region *hetero-convex* displays in the 0-ms ISI condition where convex figure CEs first emerged in response in Exp. 2 (mean: 0.66; se: 0.04), $p > 0.29$. Similarly, convex figure reports for eight-region *homo-convex* displays in the 100-ms ISI condition of Exp. 1A (mean: 0.71; se: 0.04) were statistically indistinguishable from convex figure reports for eight region *hetero-convex* displays in the 50-ms ISI condition in Exp. 2 (mean: 0.78; se: 0.04), $p > 0.29$. These results suggest that resolving the ambiguity of *homo-convex* displays adds ~50 ms to the time at which CEs are evident in convex figure reports.

Experiment 3

We have interpreted the evidence presented so far as consistent with the proposal that convex figure CEs entail recurrent processes. In Exp. 3, we used dichoptic presentations to investigate whether the relevant recurrent processes extend to the thalamus or whether they operate solely within the cortex. In dichoptic presentations, test displays and masks are presented to different eyes, as illustrated in Figure 5A. Thalamic units are monocular. The first units that respond to combined input from both eyes are in cortical area V1. Therefore, with dichoptic presentations, mask-induced activation is absent from the thalamus at least until feedback from area V1 and higher affects thalamic responses. We estimate that time minimally as the time required for area V1 to respond to a stimulus—−40–60 ms after mask onset (Lamme et al., 2002; Tapia and Beck, 2014). Hence, with dichoptic presentations of the display and mask, cortico-thalamic recurrent processing would be free of mask-induced noise until minimally 40–60 ms after mask onset (and perhaps longer if feedback originates in higher-levels than V1). Therefore, if cortico-thalamic recurrent processing plays a role in either or both convex figure CEs and ambiguity resolution, the time course of the convex figure CEs should be shifted 40–60 ms earlier than observed in Exps. 1 and 2 where the test display and the mask that followed it were presented simultaneously to both eyes (as illustrated in Figure 5B). In contrast, if only cortico-cortical recurrent processing is involved, the time course of convex figure CEs and ambiguity resolution should be the same with dichoptic presentations as with the presentation conditions used in Exps. 1 and 2 (the presentation conditions used in Exps. 1 and 2 are referred to as "monoptic presentation" conditions

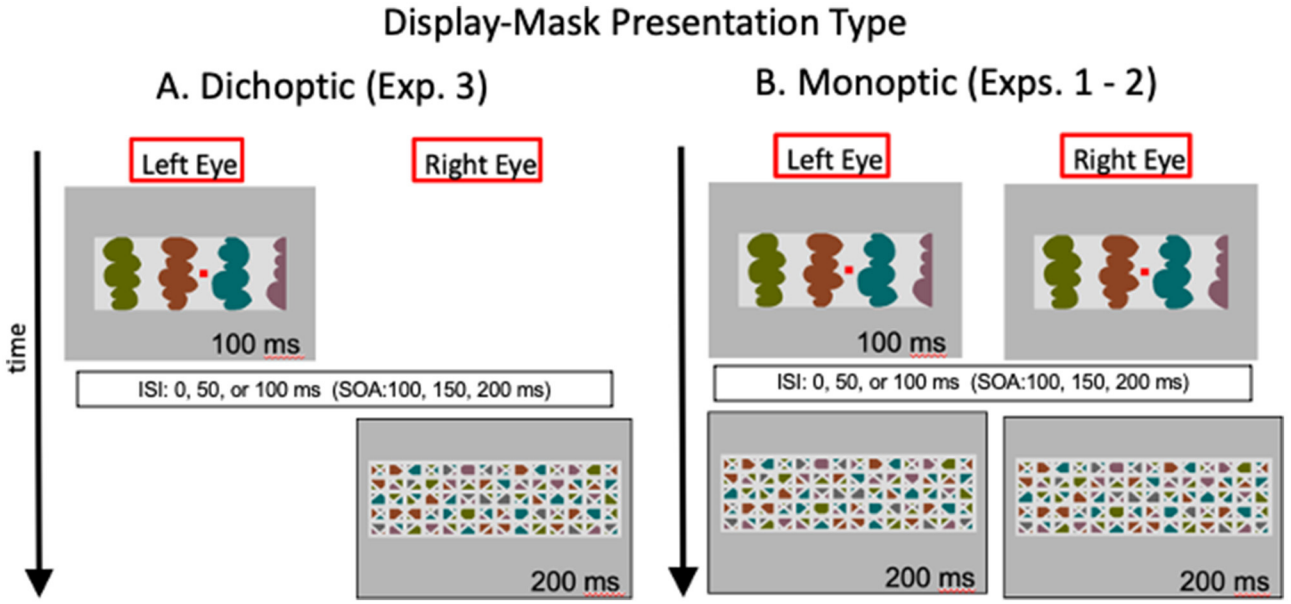


FIGURE 5
Illustration of the images in the left and right eyes as a function of time with **(A)** dichoptic presentations used in Exp. 3 and **(B)** monoptic presentations used in Exps. 1-2. ISI, interstimulus interval; SOA, stimulus onset asynchrony. *Hetero-convex* displays and masks are illustrated; *homo-convex* displays and masks like the ones in Figure 3 were also used.

because monocular as well as binocular brain regions respond to the stimuli. With monoptic presentation conditions, mask-induced activation is present as soon as activation begins in the thalamus).

We presented eight-region *homo-* and *hetero-convex* displays and masks to different groups of observers under dichoptic presentation conditions and compared the results to the results obtained for eight-region displays in Exps. 1 and 2, respectively. Only eight-region displays were used because, in previous experiments, convex figure CEs were evident in increased convex figure reports with increases in display-to-mask ISI for eight-region displays but not for two-region displays (see Figure 4 and Table 1). Moreover, the differences between *hetero-* and *homo-convex* displays were evident in convex figure reports for eight-region displays.

Participants

A total of 215 undergraduate students (147 F; 68 M) from the University of Arizona participated in Exp. 3. Of these subjects, 113 (79 F; 34 M) viewed eight-region *homo-convex* displays and their masks under dichoptic presentation conditions and 102 (68 F; 34 M) participants viewed eight-region *hetero-convex* displays intermixed with filler displays and their masks under dichoptic masking conditions. Data from 23 participants did not meet our response rate criterion; eliminating their data from the analysis left 32 participants in each of the 0, 50, and 100-ms display-to-mask ISI conditions for each display type. Assignment to ISI condition was random.

Stimuli and apparatus

A haploscope was used with a head and chin rest to present the stimuli dichoptically. In the haploscope, a pair of mirrors reflected to the left and right eyes images that were reflected to them by a second set of mirrors aimed at locations on the left and right sides of a monitor, such that each of these monitor locations were visible to one eye only (see http://www.psy.vanderbilt.edu/faculty/blake/Stereoscope/stereoscope.html). Experimental displays and masks were shown on the left and right monitor locations equally often and, hence, were presented to the left and right eyes equally often.

New sets of eight-region *homo-* and *hetero-convex* displays and filler displays were created in a size visible in the haploscope mirrors (6.45°H X 10.06°W). For each set, masks were created by cropping the masks used in Experiments 1 and 2. The masks measured 7.11°H X 12.42°W.

Procedure

Participants were seated 51.3 cm from the monitor, with distance controlled by a chinrest. Before the experimental trials, the mirrors of the haploscope were adjusted for each participant individually until the left and right-eye images of a nonius fixation cross were aligned. This procedure assured that images presented on the left and right side of the screen were aligned and centered on the fixation cross.

In each trial the test display and the mask were presented to different eyes. In half of the trials the display was presented to the left eye and the mask to the right eye; in the other half of the trials

(randomly intermixed), the display was presented to the right eye and the mask to the left eye. Participants were unaware that the images were presented to different eyes.

Participants who viewed *homo-convex* eight-region displays made their figure reports by pressing one of two vertically aligned buttons on each trial to indicate whether they perceived the black or white regions as figures. For the *hetero-convex* displays participants reported whether an elongated rectangular probe (1.45°H x 0.11°W; RGB = 255, 0, 0; luminance = 4.88 ft-L) centered vertically in either a convex or concave region to the left or right of the central edge appeared to be "on" or "off" the figure (as in Exps. 1 and 2). Peterson and Salvagio (2008) showed that these two types of response produce equivalent results. In other respects, the procedure was the same as in Exps. 1 and 2.

Data analysis

For *hetero-convex* displays, the data obtained in Exp. 3 were compared to the Exp. 2 data. For *homo-convex* displays, the data obtained in Exp. 3 were compared to the Exp. 1B data because the display widths were similar (the same results were obtained when the Exp. 3 data were compared to the Exp. 1A data).

Results and discussion

Hetero-convex displays

As can be seen in Figure 6A, convex figures were perceived in *hetero-convex* displays equally often in Exps. 2 (monoptic presentations) and 3 (dichoptic presentations) in the 0-ms and 50-ms ISI conditions. A between-experiment ANOVA showed that, for *hetero-convex* displays, convex figure reports increased as ISI increased from 0 to 50 ms, $F_{(1, 124)} = 10.506$, $p < 0.002$, $\eta^2 = 0.078$, replicating Exp. 2 (the 100-ms display-to-mask ISI condition is not included in this ANOVA because it was tested for *hetero-convex* displays only in Exp. 3; it is discussed below). Neither a main effect of presentation type nor an interaction between presentation type and ISI was observed for *hetero-convex* displays, Fs < 1. The absence of an effect of presentation type is consistent with the interpretation that cortico-cortical recurrent processes are involved in generating the convex figure reports in eight-region *hetero-convex* displays. This interpretation is not surprising inasmuch as evidence suggests that convexity is represented in cortical area V4 (Pasupathy and Connor, 1999) and that perceptual completion, a plausible mechanism linking disconnected homogeneous concave regions into a single surface, is represented in the cortex (Kourtzi and Kanwisher, 2001; Rauschenberger et al., 2006; Tang et al., 2014, 2018; Thielen et al., 2019).

Homo-convex displays

As can be seen in Figure 6B, convex figure reports for *homo-convex* displays were higher in Exp. 3 (dichoptic presentations) than in Exp. 1B (monoptic presentations) in the 0- and 50-ms display-to-mask ISI conditions, but not in the 100-ms ISI condition where previous experiments indicated convex figure CEs for masked *homo-convex* displays had reached asymptote. This pattern was shown to be statistically significant by main effects of presentation type (Exp) and ISI, $F_{(1, 124)} = 39.86$, $p < 0.001$, $\eta^2 = 0.24$ and $F_{(1, 124)} = 12.23$, $p < 0.002$, $\eta^2 = 0.09$, respectively. An interaction between presentation type and ISI was also observed, $F_{(1, 124)} = 4.37$, $p < 0.04$, $\eta^2 = 0.034$: Convex figure reports for *homo-convex* displays were statistically higher with dichoptic presentations than with monoptic presentations in both the 0-ms and 50-ms ISI conditions ($ps < 0.008$) but not the 100-ms ISI condition, $p > 0.06$, where previous evidence suggested that convex figure CEs for 100-ms displays reached asymptote. This finding suggests that cortico-thalamic feedback occurring up to 50 ms after stimulus offset plays a role in resolving the ambiguity of *homo-convex* displays. When interference from the mask in subcortical areas was removed for a period of time by presenting the mask to a different eye than the experimental display in Exp. 3, ambiguity resolution proceeded without interference and convex CEs reached asymptote in the 50-ms ISI condition, 50 ms earlier than when monoptic presentations of the display and mask were used in Exp. 1.

We next compared convex figure reports for *homo-* and *hetero-convex* displays obtained with dichoptic presentation conditions in Exp. 3 and found that they were equivalent (see Figure 6C). A 2 X 3 ANOVA with the factors of Display Type (*homo-* vs. *hetero-convex*) and ISI (0, 50, and 100 ms) revealed a main effect of ISI, $F_{(2, 186)} = 10.03$, $p < 0.001$, $\eta^2 = 0.097$ but not a main effect of Display Type, $F_{(1, 186)} = 0.497$, $p = 0.482$, nor an interaction between Display Type and ISI, $F_{(2, 186)} = 0.434$, $p = 0.648$. The finding that with dichoptic presentations the convex figure CEs emerge and reach asymptote for *homo-* and *hetero-convex* displays in the same display-to-mask ISI conditions suggests that cortical-thalamic recurrent processes involved in ambiguity resolution occur in parallel with cortico-cortical recurrent processes producing convex figure CEs. If ambiguity resolution occurred later, convex figure reports would reach asymptote in a longer ISI condition for *homo-* than *hetero-convex* displays even under dichoptic presentation conditions.

Discussion

In Exp. 3 when the test display and backward mask were presented to different eyes, thereby eliminating mask-induced noise in thalamic areas for some time, convex figure CEs emerged at the same display-to-mask ISI for *homo-* and *hetero-convex* displays. This finding contrasts with what was found with monoptic presentations of the experimental display and its mask (i.e., in Exps. 1 and 2), where convex figure CEs emerged later in time for *homo-convex* than for *hetero-convex* displays. We attributed the additional time to ambiguity resolution, suggested by Goldreich and Peterson's (2012) Bayesian observer (cf. Lass et al., 2017 for evidence consistent with this claim from tests of older participants). The results of Exp. 3 imply that ambiguity resolution involves a cortico-thalamic circuit. Moreover, our results suggest that, although feedback to the thalamus may occur even when displays are unambiguous (as in Jones et al., 2015; Poltoratski et al., 2019), it plays an essential role when ambiguity resolution is required.

Further research is necessary to determine how the ambiguity of *homo-convex* displays is resolved. One possibility is that feedback from the cortex enhances local convexity responses in the thalamus,

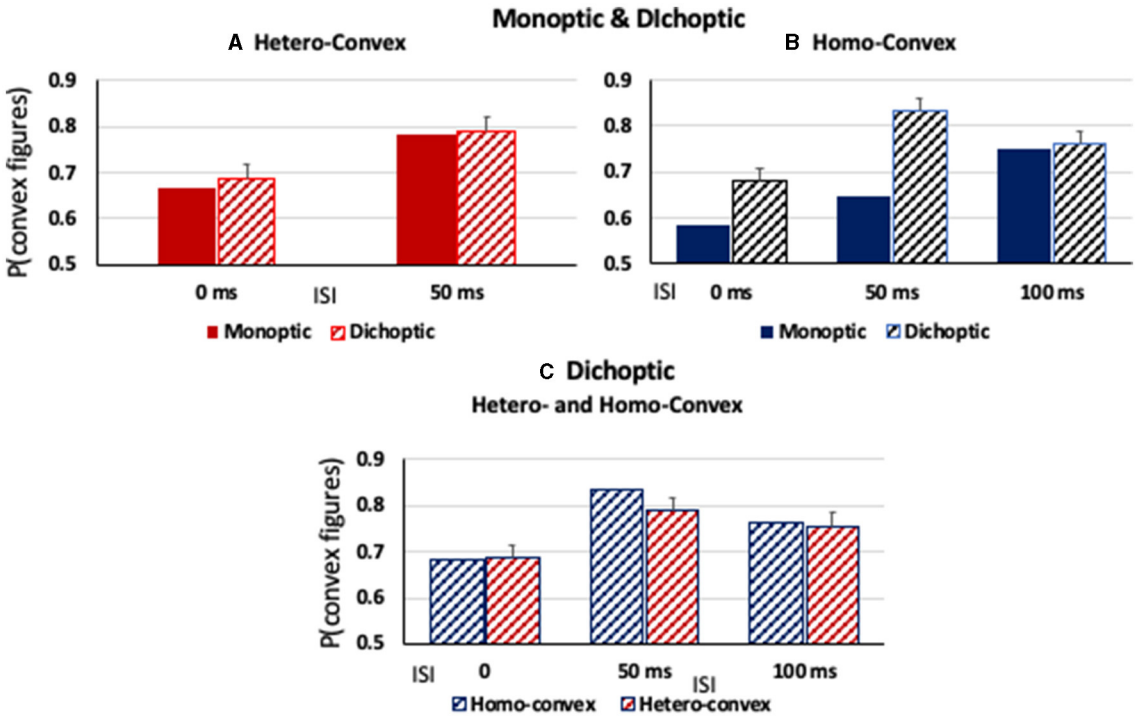


FIGURE 6
(A, B) The proportion of convex figure responses as a function of display-to-mask ISI in monoptic and dichoptic presentation conditions (solid and stiped bars, respectively) for **(A)** hetero-convex displays and **(B)** homo-convex displays. **(C)** The proportion of convex figure responses with dichoptic presentation conditions for homo- and hetero-convex displays (blue and red, respectively). Error bars represent standard errors.

thereby iteratively facilitating their transmission to higher cortical levels (cf., Jones et al., 2015; Poltoratski et al., 2019). Another possibility is that enhanced local convexity responses could bias lower-level border ownership cells (Von der Heydt, 2015) toward the convex side. Note that bias toward convexity is insufficient for convex figure CEs, however, homogeneous concave regions are necessary and evidence for convex figures increases with the number of alternating convex and homogenous concave regions (see Figure 1E). Hence, eight-region *homo-convex* displays are globally ambiguous; this ambiguity is most likely represented in regions of the cortex with receptive fields large enough to encompass eight-region displays (10–14° in the experiments presented here). Thus, it is likely that high levels of the visual hierarchy are engaged in the iterative cortico-thalamic activity that plays a role in resolving the ambiguity of *homo-convex* displays. Indeed, Sillito and Jones (2002) proposed that corticofugal feedback optimizes the thalamic contribution to global integration and segmentation.

Another possibility is that iterative cortico-thalamic activity interacts with cortical mechanisms involved in inhibitory competition between the two possible interpretations of *homo-convex* displays. Lass et al. (2017) found that older participants showed reduced or no convex figure CEs for *homo-convex* displays whereas they showed intact convex figure CEs for *hetero-convex* displays. They attributed their results to impaired suppressive mechanisms involved in inhibitory competition in older participants (cf., Betts et al., 2005, 2009; Anderson et al., 2016). There is some evidence that cortico-thalamic recurrent processing may be slower in older individuals (Walsh, 1976; Kline and Birren, 2007). Given that possibility, aging effects may instead or in addition reveal deficits in iterative cortico-thalamic processing. Using dichoptic presentations with older individuals could be informative in this regard.

Evidence indicates that the pulvinar of the thalamus is involved in attentional selection that requires distractor filtering (e.g., Snow et al., 2009; Strumpf et al., 2013). Like distractor filtering, ambiguity resolution entails a form of selection. Selecting one interpretation of an ambiguous stimulus may occur via fine-tuning cortical responses for that interpretation in one or many levels of the visual hierarchy. Ketteler et al. (2014) made a similar proposal regarding the role of cortico-thalamic recurrent processing in resolving linguistic ambiguity (cf., Mestres-Missé et al., 2017). More research is needed to determine the nature of the mechanisms initiated by cortico-thalamic feedback. Visualizing thalamic responses with high resolution fMRI is one avenue we hope to pursue in this regard. Examining perceptual organization in individuals with thalamic lesions is another.

Ambiguity resolution during perceptual organization can yield a non-reversible percept

There is no indication that *homo-convex* displays are reversible once the conflict between the object prior and the background prior for convex regions has been resolved and the best fitting interpretation has been found. Ample previous research has shown

that, given enough time without mask interference, convex figures are perceived by the vast majority of participants who view *homo-convex* displays (e.g., Koffka, 1935; Rubin, 1958; Kanizsa and Gerbino, 1976; Peterson et al., 1998; Bertamini and Lawson, 2008; Peterson and Salvagio, 2008; Barense et al., 2012; Bertamini and Wagemans, 2013; Spanò et al., 2016). Nevertheless, as suggested by our evidence, multiple interpretations are generated during perceptual organization and the best-fitting interpretation is perceived. This is exactly what is expected in a Bayesian brain – the generation of multiple interpretations for perceptual input before the best interpretation is perceived. Our results show that, for *homo-convex* displays, the processes involved in assigning figure and ground are more dynamic than assumed in traditional theories.

Symmetric figure CEs have also been reported (Mojica and Peterson, 2014). Like convexity, symmetry is an object prior, although since it requires a comparison of the two sides of a region it is necessarily more global than convexity. It also may be a weaker object prior than convexity (Kanizsa and Gerbino, 1976; Pomerantz and Kubovy, 1986; but see Mojica and Peterson, 2014). It would be interesting to examine the time course of symmetric figure CEs to investigate how symmetry interacts with the background prior and how conflict between the two priors is resolved in *homo-symmetric* displays.

Alternative interpretations

Can the results of our experiments be due to the disruption of feedforward activity in high levels of the visual hierarchy rather than to the disruption of recurrent processes? We consider that unlikely for the following reasons: First, the earlier emergence of convex figure CEs for *hetero-* than *homo-convex* displays cannot be explained by earlier high-level processing of the former than the latter. The *hetero-convex* displays are lower in contrast that the homo-convex displays. Feedforward spikes from low contrast images are delayed relative to those from higher contrast images (VanRullen and Thorpe, 2001, 2002; Wyatte et al., 2012, 2014). Therefore, based on estimates of the time for feedforward spikes to reach the cortex alone, one would expect CEs to emerge earlier in time for *homo-convex* displays than for *hetero-convex* displays. This is the opposite of what we found. Second, that convex figure CEs were no longer delayed for *homo-* relative to *hetero-convex* displays with dichoptic presentations implicates the thalamus in resolving the ambiguity of *homo-convex* displays (although the alternative interpretations may be generated in high levels, ambiguity resolution seems to require thalamic involvement). Third, as mentioned previously, differential difficulty of figure-ground decisions made in high levels cannot account for the differences between *hetero-* and *homo-*convex displays observed in Exps. 1 and 2 because those differences are not evident in Exp. 3.

Can factors other than conflict resolution account for the differences we observed between *homo-* and *hetero-convex* displays? The convex and concave regions in the former displays differ in luminance only, whereas those in the latter displays differ in both color and luminance. There is some evidence that stimuli defined by luminance differences only are processed differently from those defined by both luminance and color differences. For instance, Rivest and Cavanagh (1996) showed that borders are localized better in 2-D space when signaled by two attributes rather than one. But the conflict in our displays doesn't entail differential localization of borders in 2-D space; it involves determining whether the borders are contours of convex or concave objects. Moreover, the finding that the CEs evolve at the same time for *homo-* and *hetero-convex* displays with dichoptic presentations indicates that contour localization differences cannot account for the differences observed with monoptic presentations.

Since all conditions were manipulated between subjects rather than within subjects, might the differences between conditions be attributed to group differences rather than to the manipulated variables? Between-subjects designs were used to eliminate the influence of one condition on another. The difference between two-region and eight-region displays is critical for the CEs. Convex regions are perceived as figures much more often in eight-region than in two-region displays. We were concerned that experience with eight-region displays would contaminate convex figure reports for two-region displays, thereby reducing the difference between those two conditions. Each subject responded to many trial-unique displays within the condition in which they were tested to allow a reliable estimate of behavior in that condition. We do not believe that group differences rather than condition effects account for our results because they are replicated by different groups in different experiments (e.g., Exps. 1A and 1B; Exp. 2 monoptic results were replicated in Exp. 3 dichoptic results; and Exp. 3 hetero- and homo-convex results are not different).

It would be interesting to use a within-subjects design to test the questions addressed here and to include more fine-grained manipulation of ISI. It would be difficult, although not impossible, to present trial-unique displays in a within-subjects experiment, so the conditions would necessarily be somewhat different. Although we did not report the results in the body of the paper, we did test intermediate ISIs of 25 and 75 ms for homo-convex displays with dichoptic presentation conditions and found the results fell between the results obtained for the adjacent ISI conditions.

Conclusion

The three experiments presented here are consistent with the interpretation that recurrent cortico-thalamic processes are involved in resolving the ambiguity of eight-region *homo-convex* displays and suggest that cortico-cortical recurrent processes play a role in generating convex figure CEs.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Human Subjects Protection Program UArizona. The studies were conducted in accordance with the local legislation and institutional

requirements. The participants provided their written informed consent to participate in this study.

Author contributions

MP conceptualized the experiments, performed statistical analyses, engaged in theoretical discussions, and wrote the paper. EC created the stimuli, wrote the programs, conducted the experiments, performed statistical analysis, and engaged in theoretical discussions. All authors contributed to the article and approved the submitted version.

Funding

MP was supported by NSF (BCS 0960529) and ONR (ONR N00014-14-1-0671).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Anderson, J. A. E., Healey, M. K., Hasher, L., and Peterson, M. A. (2016). Age-related deficits in inhibition and figure-ground assignment in stationary displays. *J. Vis*. 16, 1–12. doi: 10.1167/16.7.6

Barense, M. G., Ngo, J., Hung, L., and Peterson, M. A. (2012). Interactions of memory and perception in amnesia: the figure-ground perspective. *Cerebr. Cortex* 22, 2680–2691. doi: 10.1093/cercor/bhr347

Bertamini, M., and Lawson, R. (2008). Rapid figure-ground responses to stereograms reveal an advantage for a convex foreground. *Perception* 37, 483–494. doi: 10.1068/p5728

Bertamini, M., and Wagemans, J. (2013). Processing convexity and concavity along a 2-D contour: figure-ground, structural shape, and attention. *Psychon. Bull. Rev*. 20, 191–207. doi: 10.3758/s13423-012-0347-2

Betts, L. R., Sekuler, A. B., and Bennett, P. J. (2009). Spatial characteristics of center19 surround antagonism in younger and older adults. *J. Vis.* 9, 1–15. doi: 10.1167/9.1.25

Betts, L. R., Taylor, C. P., Sekuler, A. B., and Bennett, P. J. (2005). Aging reduces center-surround antagonism in visual motion processing. *Neuron* 45, 361–366. doi: 10.1016/j.neuron.2004.12.041

Breitmeyer, B., and Ogmen, H. (2006). *Visual Masking: Time Slices Through Conscious and Unconscious Vision*. Oxford University Press.

Breitmeyer, B., and Ogmen, H. (2022). Temporal aspects of visual perception and cognition. *Oxford Res. Encycl. Psychol*. doi: 10.1093/acrefore/9780190236557.013.891

Brooks, J. L., and Palmer, S. E. (2011). Cue competition affects temporal dynamics of edge-assignment in human visual cortex. *J. Cogn. Neurosci.* 23, 631–644. doi: 10.1162/jocn.2010.21433

Bullier, J. (2001). Integrated model of visual processing. *Brain Res. Rev*. 36, 96–107. doi: 10.1016/S0165-0173(01)00085-6

Burrola, M., and Peterson, M. A. (2014). Global influences on figure assignment: the role of the border. *J. Vis.* 14, 252. doi: 10.1167/14.10.252

Craft, E., Schutze, H., Niebur, E., and von der Heydt, R. (2007). A neural model of figure–ground organization. *J. Neurophysiol.* 97, 4310–4326. doi: 10.1152/jn.00203.2007

Di Lollo, V. (2007). Iterative reentrant processing: a conceptual framework for perception and cognition (the blinding problem? No worries, mate). *Tutor. Vis. Cogn*. 2010, 9–42.

Fahrenfort, J. J., Scholte, H. S., and Lamme, V. A. F. (2007). Masking disrupts reentrant processing in human visual cortex. *J. Cogn. Neurosci.* 19, 1488–1497. doi: 10.1162/jocn.2007.19.9.1488

Gilbert, C. D., and Li, W. (2013). Top-down influences on visual processing. *Nat. Rev. Neurosci*. 14, 350–363. doi: 10.1038/nrn3476

Girard, P., Hupé, J. M., and Bullier, J. (2001). Feedforward and feedback connections between areas V1 and V2 of the monkey have similar rapid conduction velocities. *J. Neurophysiol.* 85, 1328–1331. doi: 10.1152/jn.2001.85.3.1328

Goldreich, D., and Peterson, M. A. (2012). A bayesian observer replicates convexity context effects. *Seeing Perceiv.* 25, 365–395. doi: 10.1163/187847612X634445

Grossberg, S., and Mingolla, E. (1985). Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading. *Psychol. Rev*. 92, 173. doi: 10.1037/0033-295X.92.2.173

Guttman, S., Sekuler, A. B., and Kellman, P. J. (2003). Temporal variations in visual completion: a reflection of spatial limits.? *JEP:HPP* 29, 1211–1227. doi: 10.1037/0096-1523.29.6.1211

Hochberg, J. (1971). "Perception I: color and shape," in *Woodworth and Schlossberg's Experimental Psychology, 3rd Edn*, eds J. W. Kling and L. A. Riggs (New York, NY: Holt, Rinehart, & Winston), 395–474.

Hoffman, D. D., and Singh, M. (1997). Salience of visual parts. *Cognition* 63, 29–78. doi: 10.1016/S0010-0277(96)00791-3

Jehee, J. F., Lamme, V. A., and Roelfsema, P. R. (2007). Boundary assignment in a recurrent network architecture. *Vis. Res.* 47, 1153–1165. doi: 10.1016/j.visres.2006.12.018

Jones, H. E., Andolina, I. M., Shipp, S. D., Adams, D. L., Cudeiro, J., Salt, T. E., et al. (2015). Figure-ground modulation in awake primate thalamus. *Proc. Natl. Acad. Sci. U. S. A.* 112, 7085–7090. doi: 10.1073/pnas.1405162112

Kanizsa, G., and Gerbino, W. (1976). "Convexity and symmetry in figure-ground organization," in *Vision and Artifact*, ed M. Henle (New York, NY: Springer), 1–57.

Kellman, P. J., and Shipley, T. F. (1991). A theory of visual interpolation in object perception. *Cogn. Psychol.* 23, 141–221. doi: 10.1016/0010-0285(91)90009-D

Kelly, F., and Grossberg, S. (2000). Neural dynamics of 3-D surface perception: figure-ground separation and lightness perception. *Percept. Psychophys*. 62, 1596–1618. doi: 10.3758/BF03212158

Ketteler, S., Ketteler, D., Vohn, R., Kastrau, F., Schulz, J. B., Reetz, K., et al. (2014). The processing of lexical ambiguity in healthy ageing and Parkinson? s disease: role of cortico subcortical networks. *Brain Res.* 1581, 51–63. doi: 10.1016/j.brainres.2014.06.030

Kline, D. W., and Birren, J. E. (2007). Age differences in backward dichoptic masking. *Exp. Aging Res.* 1, 17–25. doi: 10.1080/03610737508257943

Koffka, K. (1935). *Principles of Gestalt Psychology*. London: Harcourt Brace & World, Inc.

Kourtzi, Z., and Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital complex. *Science* 293, 1506–1509. doi: 10.1126/science.1061133

Kreiman, G., and Serre, T. (2020). Beyond the feedforward sweep: feedback computations in the visual cortex. *Ann. N. Y. Acad. Sci*. 1464, 222–241. doi: 10.1111/nyas.14320

Lamme, V. A. (1995). The neurophysiology of figure-ground segregation in primary visual cortex. *J. Neurosci.* 15, 1605–1615. doi: 10.1523/JNEUROSCI.15-02-01605.1995

Lamme, V. A., and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* 23, 571–579. doi: 10.1016/S0166-2236(00)01657-X

Lamme, V. A., Zipser, K., and Spekreijse, H. (2002). Masking interrupts figure-ground signals in V1. *J. Cogn. Neurosci.* 14, 1044–1053. doi: 10.1162/089892902320474490

Lass, J. W., Bennett, P. J., Peterson, M. A., and Sekuler, A. B. (2017). Effects of aging on figure-ground perception: convexity context effects and competition resolution. *J. Vis.* 17, 1–16. doi: 10.1167/17.2.15

Macknik, S. L., and Martinez-Conde, S. (2007). The role of feedback in visual masking and visual processing. *Adv. Cogn. Psychol.* 3, 125–153. doi: 10.2478/v10053-008-0020-5

Mestres-Missé, A., Trampel, R., Turner, R., and Kotz, S. A. (2017). Uncertainty and expectancy deviations require cortico-subcortical cooperation. *NeuroImage* 144, 23–34. doi: 10.1016/j.neuroimage.2016.05.069

Mojica, A. J., and Peterson, M. A. (2014). Display-wide influences on figure-ground perception: the case of symmetry. *Attent. Percept. Psychophys.* 76, 1069–1084. doi: 10.3758/s13414-014-0646-y

O'Shea, R. P., Blackburn, S. G., and Ono, H. (1994). Contrast as a depth cue. *Vis. Res.* 34, 1595–1604. doi: 10.1016/0042-6989(94)90116-3

Pasupathy, A., and Connor, C. E. (1999). Responses to contour features in macaque area V4. *J. Neurophysiol.* 82, 2490–2502. doi: 10.1152/jn.1999.82.5.2490

Peterson, M. A., and Enns, J. T. (2005). The edge complex: implicit perceptual memory for cross-edge competition leading to figure assignment. *Percept. Psychophys.* 4, 727–740. doi: 10.3758/BF03193528

Peterson, M. A., Gerhardstein, P. C., Mennemeier, M., and Rapcsak, S. Z. (1998). Object-centered attentional biases and object recognition contributions to scene segmentation in left- and right- hemisphere-damaged patients. *Psychobiology* 26, 557–570. doi: 10.3758/BF03330622

Peterson, M. A., and Lampignano, D. L. (2003). Implicit memory for novel figure-ground displays includes a history of border competition. *J. Exp. Psychol.* 29, 808–822. doi: 10.1037/0096-1523.29.4.808

Peterson, M. A., Mojica, A. J., Salvagio, E., and Kimchi, R. (2017). Figural properties are prioritized for search under conditions of uncertainty: setting boundary conditions on claims that figures automatically attract attention. *Attent. Percept. Psychophys.* 79, 180–199. doi: 10.3758/s13414-016-1223-3

Peterson, M. A., and Salvagio, E. (2008). Inhibitory competition in figure-ground perception: context and convexity. *J. Vis.* 8, 1–13. doi: 10.1167/8.16.4

Poltoratski, S., Maier, A., Newton, A. T., and Tong, F. (2019). Figure-ground modulation in the human lateral geniculate nucleus is distinguishable from top-down attention. *Curr. Biol.* 29, 2051–2057. doi: 10.1016/j.cub.2019.04.068

Pomerantz, J. R., and Kubovy, M. (1986). "Theoretical approaches to perceptual organization: simplicity and likelihood principles," in *Handbook of Perception and Performance, Volume II: Cognitive Processes and Performance*, eds K. R. Boff, L. Kaufman, and J. P. Thomas (New York, NY: John Wiley & Sons), 1–46.

Rauschenberger, R., Liu, T., Slotnick, S. D., and Yantis, S. (2006). Temporally unfolding neural representation of pictorial occlusion. *Psychol. Sci.* 17, 358–364. doi: 10.1111/j.1467-9280.2006.01711.x

Ringach, D. L., and Shapley, R. (1996).Spatial and temporal properties of illusory contours and amodal boundary completion. *Vis. Res.* 36, 3037–3050. doi: 10.1016/0042-6989(96)00062-4

Rivest, J., and Cavanagh, P. (1996). Localizing contours defined by more than one attribute. *Vis. Res.* 36, 53–66. doi: 10.1016/0042-6989(95)00056-6

Roelfsema, P. R. (2006). Cortical algorithms for perceptual grouping. *Annu. Rev. Neurosci.* 29, 203–227. doi: 10.1146/annurev.neuro.29.051605.112939

Rubin, E. (1958). "Figure and ground," in *Readings in Perception*, eds D. Beardslee and M. Wertheimer (Princeton, NJ: Van Nostrand), 35–101.

Sekuler, A. B., and Palmer, S. E. (1992). Perception of partly occluded objects: a microgenetic analysis. *J. Exp. Psychol.* 121, 95–111. doi: 10.1037/0096-3445.121.1.95

Self, M. W., Jeurissen, D., van Ham, A. F., van Vugt, B., Poort, J., and Roelfsema, P. R. (2019). The segmentation of proto-objects in the monkey primary visual cortex. *Curr. Biol.* 29, 1019–1029. doi: 10.1016/j.cub.2019.02.016

Sillito, A. M., and Jones, H. E. (2002). Corticothalamic interactions in the transfer of visual information. *Philos. Trans. Royal Soc. Lond. Ser. B* 357, 1739–1752. doi: 10.1098/rstb.2002.1170

Snow, J. C., Allen, H. A., Rafal, R. D., and Humphreys, G. W. (2009). Impaired attentional selection following lesions to human pulvinar: evidence for homology between human and monkey. *Proc. Natl. Acad. Sci. U. S. A.* 106, 4054–4059. doi: 10.1073/pnas.0810086106

Spanò, G., Peterson, M. A., Nadel, L., Rhoads, C., and Edgin, J. O. (2016). Seeing can be remembering: interactions between memory and perception in typical and atypical development. *Clin. Psychol. Sci.* 4, 254–271. doi: 10.1177/2167702615590997

Strumpf, H., Mangun, G. R., Boehler, C. N., Stoppel, C., Schoenfeld, M. A., Heinze, H. J., et al. (2013). The role of the pulvinar in distractor processing and visual search. *Hum. Brain Map.* 34, 1115–1132. doi: 10.1002/hbm.21496

Tang, H., Buia, C., Madhavan, R., Crone, N. E., Madsen, J. R., Anderson, W., et al. (2014). Spatiotemporal dynamics underlying object completion in human ventral visual cortex. *Neuron* 83, 736–748. doi: 10.1016/j.neuron.2014.06.017

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., et al. (2018). Recurrent computations for visual pattern completion. *Proc. Natl. Acad. Sci. U. S. A.* 115, 8835–8840. doi: 10.1073/pnas.1719397115

Tapia, E., and Beck, D. M. (2014). Probing feedforward and feedback contributions to awareness with visual masking and transcranial magnetic stimulation. *Front. Psychol.* 5, 1173. doi: 10.3389/fpsyg.2014.01173

Thielen, J., Bosch, S. E., van Leeuwen, T. M., van Gerven, M. A. J., and van Lier, R. (2019). Neuroimaging findings on amodal completion: a review. *i-Perception* 10, 041669519840047 doi: 10.1177/2041669519840047

VanRullen, R., and Thorpe, S. J. (2001). The time course of visual processing: from early perception to decision-making. *J. Cogn. Neurosci.* 13, 454–461. doi: 10.1162/08989290152001880

VanRullen, R., and Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vis. Res.* 42, 2593–2615. doi: 10.1016/S0042-6989(02)00298-5

Von der Heydt, R. (2015). Figure–ground organization and the emergence of proto-objects in the visual cortex. *Front. Psychol.* 6, 1695. doi: 10.3389/fpsyg.2015.01695

Walsh, D. A. (1976). Age differences in central perceptual processing: a dichoptic backward masking investigation. *J. Gerontol.* 31, 178–185. doi: 10.1093/geronj/31.2.178

Wyatte, D., Curran, T., and O'Reilly, R. (2012). The limits of feedforward vision: recurrent processing promotes robust object recognition when objects are degraded. *J. Cogn. Neurosci.* 24, 2248–2261. doi: 10.1162/jocn_a_00282

Wyatte, D., Jilk, D. J., and O'Reilly, R. C. (2014). Early recurrent feedback facilitates visual object recognition under challenging conditions. *Front. Psychol.* 5, 674. doi: 10.3389/fpsyg.2014.00674

Yin, C., Kellman, P. J., and Shipley, T. F. (1997). Surface completion complements boundary interpolation in the visual integration of partly occluded objects. *Perception* 26, 1459–1479. doi: 10.1068/p261459

Yin, C., Kellman, P. J., and Shipley, T. F. (2000). Surface integration influences depth discrimination. *Vis. Res.* 40, 1969–1978. doi: 10.1016/S0042-6989(00)00047-X

Zhou, H., Friedman, H. S., and Von Der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *J. Neurosci.* 20, 6594–6611. doi: 10.1523/JNEUROSCI.20-17-06594.2000

Zipser, K., Lamme, V. A. F., and Schiller, P. H. (1996). Contextual modulation in primary visual cortex. *J. Neurosci.* 16, 7376–7389. doi: 10.1523/JNEUROSCI.16-22-07376.1996

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 15 November 2023
DOI 10.3389/fpsyg.2023.1198691

OPEN ACCESS

EDITED BY
James Elder,
York University, Canada

REVIEWED BY
Yingle Fan,
Hangzhou Dianzi University, China
Udo Ernst,
University of Bremen, Germany

*CORRESPONDENCE
Doreen Hii
✉ doreen.hii@uci.edu

RECEIVED 01 April 2023
ACCEPTED 27 October 2023
PUBLISHED 15 November 2023

CITATION
Hii D and Pizlo Z (2023) Combining contour and region for closed boundary extraction of a shape. *Front. Psychol.* 14:1198691. doi: 10.3389/fpsyg.2023.1198691

Combining contour and region for closed boundary extraction of a shape

Doreen Hii* and Zygmunt Pizlo

Visual Perception Laboratory, Department of Cognitive Sciences, University of California, Irvine, Irvine, CA, United States

This study explored human ability to extract closed boundary of a target shape in the presence of noise using spatially global operations. Specifically, we investigated the contributions of contour-based processing using line edges and region-based processing using color, as well as their interaction. Performance of the subjects was reliable when the fixation was inside the shape, and it was much less reliable when the fixation was outside. With fixation inside the shape, performance was higher when both contour and color information were present compared to when only one of them was present. We propose a biologically-inspired model to emulate human boundary extraction. The model solves the shortest (least-cost) path in the log-polar representation, a representation which is a good approximation to the mapping from the retina to the visual cortex. Boundary extraction was framed as a global optimization problem with the costs of connections calculated using four features: distance of interpolation, turning angle, color similarity and color contrast. This model was tested on some of the conditions that were used in the psychophysical experiment and its performance was similar to the performance of subjects.

KEYWORDS

boundary extraction, contour, color, log-polar representation, Dijkstra algorithm

1 Introduction

Boundary extraction involves identifying and connecting a set of visual elements such as line edges to form the boundary of an object. Boundary extraction is one of the first, if not the very first operations that the human visual system performs. Given the vast amount of information present in any visual scene, the computations performed by the human visual system must be robust. Specifically, the visual system must be able to ignore irrelevant information and it should be insensitive to errors which could occur during edge detection. Top row in Figure 1 illustrates how our stimuli looked. The egg-like shape in the left panel is easier to see than the one in the right panel. This is because the orientations of the edges defining the boundary of the shape were perfect in the left panel while the orientations were randomly perturbed in the right panel. When orientations of edges form a smooth contour, local interpolation could extract the target boundary (Bottom left of Figure 1). Local interpolation would fail when jitter level is high (Figure 1 Bottom right), highlighting the need for global operations in extracting the boundary. This study investigated human global operations in boundary extraction with a focus on the integration of contour and region information.

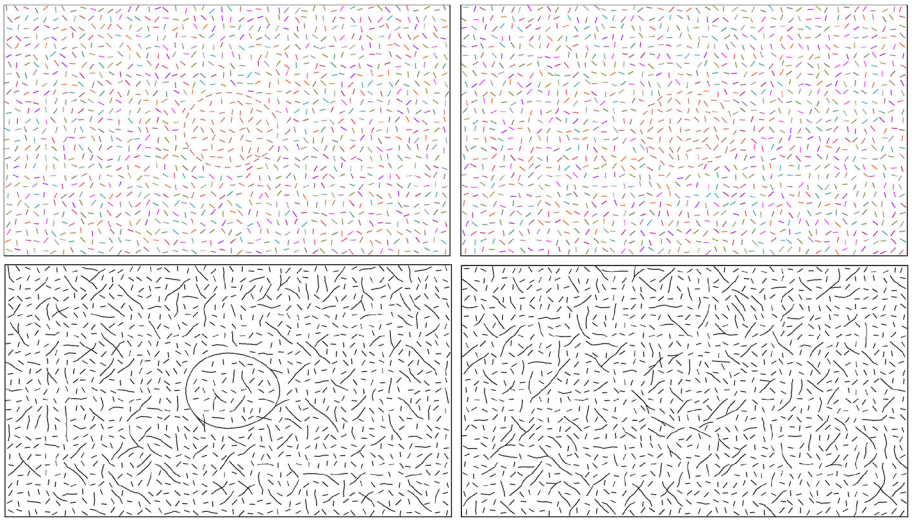


FIGURE 1
Top row: Examples of stimuli used in our experiment. Contrast was reversed in the actual experiment. **Top left**: Orientation jitter of the edges in the boundary of the shape is zero. **Top right**: The orientation jitter is 20°. **Bottom row**: Outputs of a local interpolation algorithm that connected neighboring edges when turning angle was ≤40°. **Bottom left**: This local interpolation was able to extract the boundary of the target egg. **Bottom right**: Local interpolation failed to extract the boundary of the egg, demonstrating the need for global operations in extracting the boundaries in our experiments.

In general, the human visual system may use two types of information to accomplish boundary extraction: contour information such as that encoded in edges and region information such as that encoded in color (Grossberg and Mingolla, 1985). Since humans are able to extract boundaries in isochromatic and isoluminant stimuli, the visual system can use either type of information to independently arrive at a boundary solution. Nonetheless, since both contour and color are available in most cases of everyday life, they are jointly encoded in all areas of the lower (V1-V4) and higher regions (lateral and ventral occipitotemporal) of the ventral visual stream (Taylor and Xu, 2022).

Perceptually, when both contour and color are available, the effect of contours seems to dominate over that of color. The McCullough Effect is such an example where the afterimage after viewing two regions with different line orientations and colors depends on the orientation of the lines (Tyler and Solomon, 2019). Further evidence was provided by Vergeer et al. (2015) who demonstrated malleable color percept: different placements of edges created different color percepts, and color inside the shape boundary was always consistent. In fact, previous research has suggested that shape from contour is the fastest cue available to the visual system (Elder, 2018) and is the necessary prerequisite before color-based processing (Moutoussis, 2015). While contour information may adequately suggest a boundary in many cases, color may improve performance when edges are noisy. Hansen and Gegenfurtner (2009) have shown that the two pieces of information are not redundant copies of each other. For example, color is less sensitive to changes in shading or lighting. Therefore, changes in color better indicate a change in material which could in turn suggest the presence a new object (Moutoussis, 2015). Moreover, Taylor and Xu (2023) have found that the cortical areas in the ventral visual pathway could increase the relative coding strength for color depending on the type of stimuli being presented (simple orientation or complex tessellation patterns).

In this study, we (1) performed psychophysical experiments investigating the integration of contour and color using conditions where spatially global operations are required, as well as (2) developed a computational model to emulate human performance. Results from our psychophysical experiments showed that while contour and color information could be utilized in isolation, performance was highest and most robust when they worked in conjunction. Moreover, our results also suggested that contour and color could be integrated only with foveal viewing. With peripheral viewing, performance dropped from ceiling to chance when the orientation jitter increased from 0° to 20°. The failure of boundary discrimination with peripheral viewing could not be explained by a decrease in visual resolution.

Following the study by Kwon et al. (2016), our model uses the log-polar representation of the retinal image. It is known that a log-polar map is a good approximation of the transformation from the retinal image to the early visual areas in the cortex (starting with V1), the first areas responsible for extraction of contours and boundaries (Schwartz, 1977). The log-polar transformation preserves spatially local relations, which means that local neighborhoods in the retinal image are mapped into local neighborhoods in the visual cortex. It follows that there may only be small differences in how computational models work when using the retinal versus the log-polar representation. However, spatially global computations may look very different in the retinal (Cartesian) coordinate system versus in the cortical (log-polar) coordinate system because the log-polar mapping

distorts spatially global relations[1]. Not only so, we believe that the concept of log-polar is tightly related to other known visual architectures such as the multiresolution / multiscale pyramid (Rosenfeld and Thurston, 1971; Tanimoto and Pavlidis, 1975).[2] In the present study we emphasized *spatially global computations that result in a closed boundary of a 2D region on the retina*. We further argue that the log-polar representation is essential in guaranteeing a closed boundary solution when spatially local computations are insufficient (such as that in Figure 1 right column).

We substantially elaborated the previous model proposed by Kwon et al. (2016). Similar to Kwon et al. (2016), the current study focused on the conditions which are perceptually difficult, namely when orientation jitter was added to remove local contour cues. Therefore, the target shape in our stimuli would be difficult or even impossible to detect for a computational model that uses only local operations (see Figure 1). Additionally, we showed that subjects' performance improved when color information was made available. Thus, two requirements were placed on the computational model: (1) the model must perform global operations to be able to accurately produce a closed boundary, and (2) the model must be able to combine both contour and color information. Our proposed model guarantees closure and implements five other Gestalt principles including proximity, good continuation, convexity, color similarity and dissimilarity. This model was tested on some of the conditions on which the subjects were tested, and its performance was similar to the performance of the subjects. We additionally tested the model on a small set of real images and demonstrated promising results. We want to point out that the current model is not intended as the complete theory. Instead, it is the first attempt in capturing the role of contour closure, proximity, good continuation, convexity, color similarity and dissimilarity in the log-polar representation.

2 Psychophysical experiment

We extended the experiments reported in Kwon et al. (2016) where the authors measured the role of contour in boundary extraction using black-and-white line drawings. We first replicated their main result, and then performed a 2 × 2 factorial experiment involving two levels of orientation jitter applied to edges and two levels of background colors. We expected that the addition of color information would improve the performance of boundary extraction.

1 For example, consider a circle with its center coinciding with the center of the retina. Inscribe a square into the circle. The perimeter of the circle is longer than that of the inscribed square in the Cartesian coordinates representing the retina. This relationship is flipped in the log-polar space. The circle maps into a straight line, while the square maps into a four-cornered curve (see Figure 13 in Kwon et al., 2016). It follows that the straight line representing the circle is shorter than the curve representing the square.

2 Multiresolution pyramid adjusts the resolution of operations according to the scale of the object on the retinal image The log-polar architecture is scale invariant such that retinal shapes that differ in sizes are processed by the same number of neurons. Thus, the problem of adjusting scale and resolution of processing is solved naturally by a log-polar transformation.

We followed the procedure described in Experiment 3 of Kwon et al. (2016) to test boundary extraction using the fragmented boundary of an egg-like shape embedded in noise edges. The subject's task was to indicate if the pointy side of the egg was oriented to the left or to the right. This task required extraction of the entire boundary of the shape.

2.1 Methods

2.1.1 Stimuli

The stimulus consisted of boundary edges belonging to a target shape embedded in noise. In the current experiment, a stimulus canvas of size of [1,920 × 1,080 pixels] was used. To fill the canvas with noise edges, the canvas was divided into [48 × 27] square grids, each with size [40 × 40 pixels]. A noise edge with random orientation was added to each grid, with the center of the edge coinciding with the center of the grid. The edge was allowed to occupy the central 60% of its grid to prevent coincidental connections of neighboring edges.

The target shape was an egg created by distorting an ellipse (Kozma-Wiebe et al., 2006), using the following formula:

$$\frac{x^2}{5^2} \times \frac{1}{1 \pm kx} + \frac{y^2}{4^2} = 1$$

where k is the distortion coefficient such that a larger k produces an egg with a more obvious pointy side and makes the discrimination task easier. For three of our four subjects (S1-S3), we used $k = 0.04$, which was the same value used in the previous study (Kwon et al., 2016). Subject S4 was tested with $k = 0.08$. The rectangle circumscribed on the egg was 450 × 360 pixels. The horizontal radius (225 pixels) corresponded to a visual angle of 6.66° when viewed from a distance of 60cm. The continuous, smooth egg boundary was fragmented into straight line segments of similar lengths as the noise edges. Every other egg boundary edge was erased, to produce a support ratio of 0.5. The center of the egg was shifted away from the fixation cross in a random direction, with the maximum shift being 50% of the minor radius of the egg.

Then, orientation noise was added to the edges belonging to the egg boundary. Two levels of orientation jitter were used: 20° and 180°. We followed the convention in Kwon et al. (2016), where an average orientation jitter of 20° referred to random rotation of a boundary edge sampled from either [-25°, -15°] or [+15°, +25°]. Orientation jitter of 20° was chosen because contour smoothness was reported by Kwon et al. (2016) to be ineffective for local interpolation. Therefore, the experiment with this level of jitter would likely measure spatially global operations applied to the entire closed contour at once. Sensitivity to jitter was directly tested in our control experiment (Section 4.3.1). Moreover, local interpolation based on smoothness failed to extract boundary for jitter level of 20° (Figure 1 bottom right). The jitter 180° condition changed the orientation of the edge by an angle between -180° and +180°. This implied that the orientation of the contour fragments of the egg conveyed no information about the boundary of the egg. Once the egg edges were prepared, they were added to the canvas. Noise edges were removed if necessary to prevent overlapping of edges.

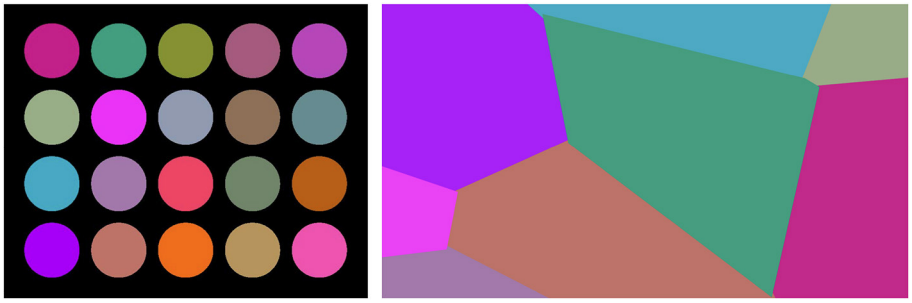


FIGURE 2
Left: Custom colormap of 20 colors used in this experiment. **Right**: An example of a Worley color pattern created by coloring eight Voronoi partitions.

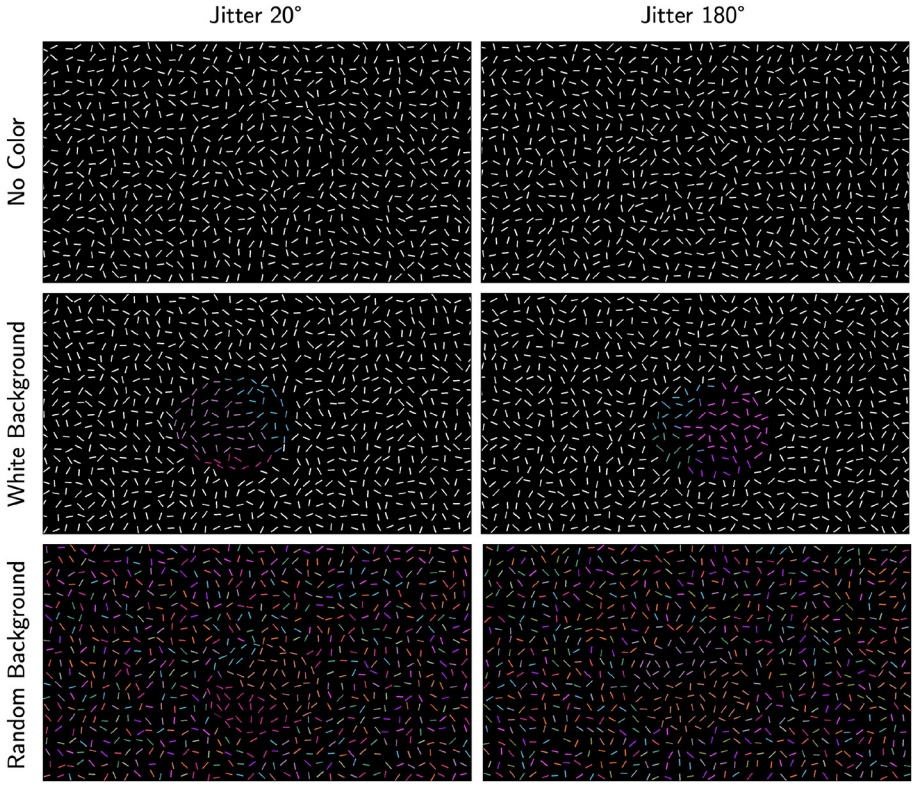


FIGURE 3
Examples of the stimuli. The **left column** shows examples with jitter 20° and the **right column** shows jitter 180°. The **first row** shows the conditions with no color, the **second** and **third rows** show examples with Worley color pattern added to all edges inside the egg (including the boundary edges). In the **second row**, edges outside the egg are white (white-background) whereas they have random color (random-background) in the **third row**.

At this point, the stimulus canvas consisted of grid-like noise edges and fragmented egg edges. All edges had thickness of one pixel. To better conceal the egg, positional jitter was added to noise edges by moving their centers by a random amount in the range [-10, +10] pixels and in a random direction, with the constraint that no edges overlapped. The resulting stimulus consisted of white edges on a black background. Examples of this stimulus are shown in Figure 3, first row.

2.1.1.1 Adding colors to edges

The edges inside the egg (including boundary edges) were colored as follows. In each trial, a new, randomly generated Worley noise pattern was used. Worley noise is a popular texture generation method to simulate real world patterns (Worley, 1996). In this study, we generated Worley color pattern by performing Voronoi partition using a set of five to nine seed points randomly positioned on the canvas. Each Voronoi partition was given a different color to generate a Worley color pattern. To make sure that the color regions were clearly visible against a black background and against the white noise edges, we constructed a custom colormap of 20 colors after restricting the categorical colors from Colorcet to only those with brightness values within the range of 40–60 (out of 100) (Bednar et al., 2020). The colors used in our experiments are shown in Figure 2, along with an example of the

Worley color pattern. Based on the position of the egg, noise edges inside the egg boundary (including the edges of the boundary) took on the colors as defined by the Worley color pattern. The number of color regions varied depending on the sizes of the Voronoi partitions.

We used two color conditions in the background. In one condition, all noise edges in the background were uniformly white (white-background), and in the other condition, each noise edge in the background was assigned a random color from the custom color map (random-background). Examples from both conditions are shown in the second and third rows of Figure 3 respectively. We expected the white-background condition (second row) to be easier than the random-background condition (third row). Note that our stimuli with color looked like watercolor illusions, namely illusory colors spread in between the blank space of edges and filled in the region (Pinna et al., 2001).

We randomized all parameters irrelevant to the manipulated variables. So for each trial, orientation and position jitter of noise edges were sampled randomly; the fragmentation of the egg had a random starting point; each edge belonging to the egg boundary had a random orientation jitter added; a new position for egg center was selected; a new Voronoi partition was generated; random color was sampled to fill each partition when creating a Worley pattern; and if applicable, the colors for noise edges outside the egg were sampled randomly from the colormap.

2.1.2 Experimental conditions

Example stimulus illustrating each of the six experimental conditions is presented in Figure 3. To improve visibility of these examples, edges are drawn with thicker lines and with a lower density of edges relative to the size of the stimulus. So, these images are not copies of our stimuli. Nevertheless, they illustrate the conditions well. The actual stimuli used in the experiment are publicly available. In Figure 3, we show right-pointing eggs in the first and third rows and left-pointing eggs in the second row.

2.1.3 Subjects

Four subjects were tested: Subject S1, who received an extensive practice before data collection; Subject S2; and two naive subjects, Subjects S3 and S4. In the main experiment, three subjects were tested with distortion coefficient $k = 0.04$ and Subject S4 was tested with a larger distortion (distortion coefficient, $k = 0.08$) to make sure that performance in most conditions was well above chance. All subjects had normal or corrected to normal vision.

2.1.4 Procedure

Signal detection experiment was used. Each session consisted of two hundred left-pointing eggs and two hundred right-pointing eggs presented in random order. Each session began with 40 warm-up trials before the 400 experimental trials. The experiments were performed in a well-lit room. Subjects viewed the stimuli with both eyes from a distance of 60cm using a chin-forehead rest. The monitor had a 60Hz refresh rate, and the measured chromaticity coordinates of the RGB primary colors and luminance values for the white point are summarized in Table 1. A trial began by displaying the fixation cross at the center of the monitor. Subjects pressed a key to advance when they were ready. A blank screen was shown for 100ms followed by the stimulus that was shown for 100ms. After that, the blank screen was shown until the subject responded by pressing "Q" if the egg pointed to the left or "P" if it pointed to the right. A beep was sounded after an incorrect response. This sequence was repeated until all 400 trials were completed. Subjects were given as much time as needed to familiarize with the task. Subjects completed one practice session before the actual data collection.

TABLE 1 Chromaticity coordinates of the RGB and luminance values of the monitor.

	x	y	Y(cd/m^2)
R	0.64	0.35	59.7
G	0.32	0.60	231
B	0.14	0.06	21.9
W	0.312	0.344	314

Subjects were first tested with jitter 20° and no-color condition (Figure 3 top left) to allow for an estimate of their distortion coefficient, k. Subjects S1 and S2 were also tested with the jitter 180° and no-color (Figure 3 top right) to verify that performance in this condition was at chance. The other two subjects (S3 and S4) were not tested in the jitter 180° and no-color condition. After that, each subject completed the four main experimental conditions in random order.

2.2 Results

Figure 4 shows the results from individual subjects. Subjects' performance was evaluated using the discriminability measure d' of signal detection. To estimate d' for a two-alternative-forced-choice (2AFC) task, one of the two stimuli (say, egg pointing to the left) can be assigned as "noise" and the other as "signal plus noise". This way, hit and false alarm rates can be computed and used to estimate d' by subtracting the Z-score of false alarms from the Z-score of hits. A higher d' represents better performance and a d' of zero indicates chance performance. Reliability of d' for each subject was estimated using the standard error of d' as described by Macmillan and Creelman (2004) (p. 325).

When no color was used, performance of the subjects was reliable with jitter 20°. Specifically, all four subjects achieved d' between 0.5 and 1.5 in this condition. In contrast, jitter 180° led to chance performance. This was expected, so only two subjects (S1 and S2) were tested in jitter 180° and no-color condition.

Next, we will describe the four conditions in which the color of edges inside the egg was different from the color of edges outside the egg. For jitter 20°, performance was equally good when the background edges were white and when the background edges had random color. At the same time, performance in these two conditions was clearly better than performance in the no-color condition. In three of the four subjects, this improvement was by a factor of 2 or more. A different pattern of results was observed

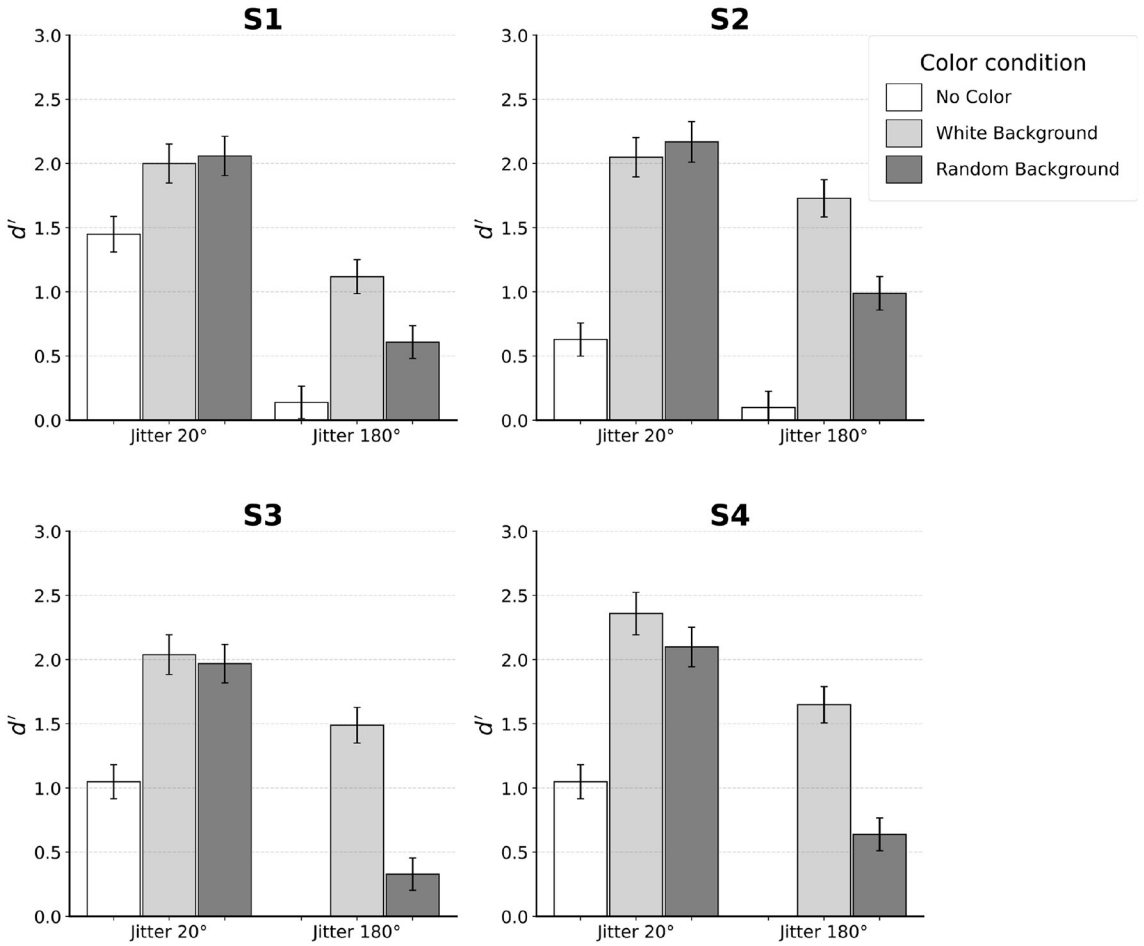


FIGURE 4
Results for each individual subject. Error bars indicate SE. Subjects S1–S3 performed the experiment with distortion of $k = 0.04$. Subject S4 was tested with $k = 0.08$.

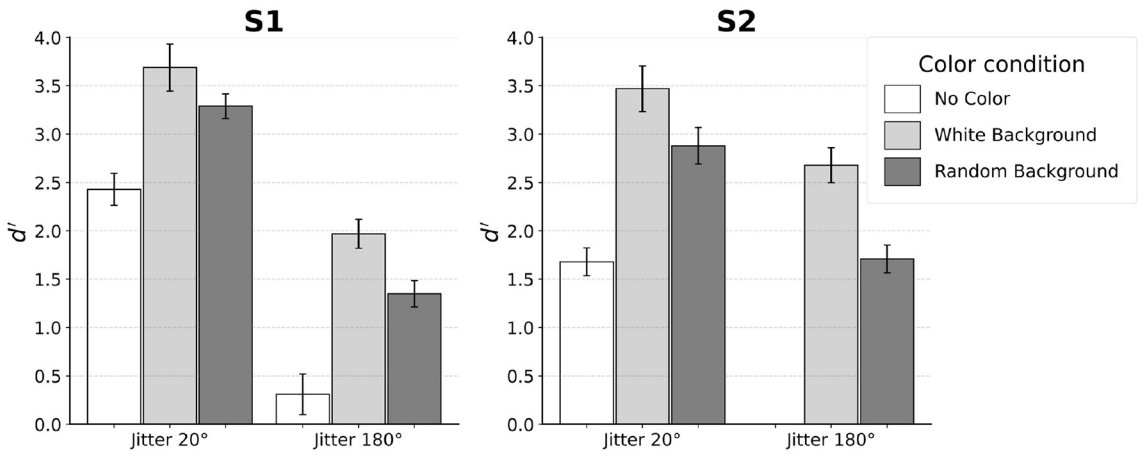


FIGURE 5
Performance with larger shape distortion ($k = 0.08$). Performance improved, but the pattern of results is the same as with $k = 0.04$.

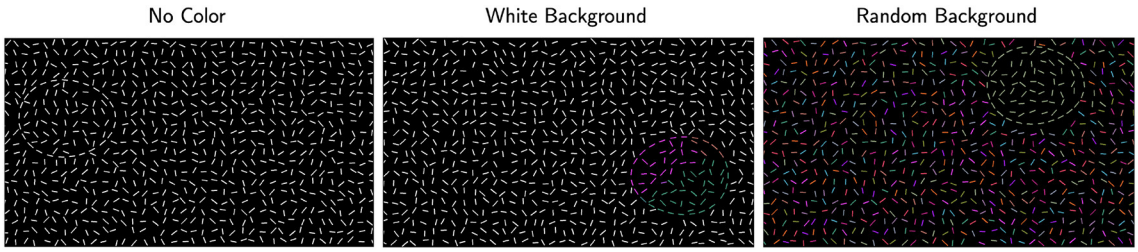


FIGURE 6
Examples of stimuli with jitter 0° for peripheral viewing. The three color conditions are shown in separate columns: no-color condition where all edges were white, white-background condition where Worley color pattern was added inside the egg (including boundary edges), and random-background condition where Worley color pattern was added inside the egg and edges in the background had random color.

for jitter 180°. Unlike the chance performance where no color was used, adding color led to performance that was above chance, especially when background was uniformly white. For background with random color, performance was lower by a factor of two on average compared to the condition where background color was white.

2.3 Discussion

We replicated the results of Kwon et al. (2016) using the jitter 20° and no-color condition by showing that subjects could reliably perform boundary extraction even with this level of jitter. Jitter 180° and no-color produced chance performance, as expected. This indicated the effectiveness of the noise edges in concealing the egg, so that no confounding cue was available for subjects to complete boundary extraction. Since no color was present in this pair of conditions, we expected performance to rely solely on contour-based processing. While contour-based processing tolerated an orientation jitter of 20°, completely randomizing orientations in the jitter 180° condition made contour integration ineffective. Thus, boundary extraction could not occur. When color information was made available, color-based processing was recruited to improve performance. Since contour-based processing was already recruited in the jitter 20° conditions, adding color reflected the joint operation of contour- and color-based processing. In the jitter 180° conditions, color was the only cue that could lead to contour extraction.

In general, adding color improved performance. However, there was a difference in the magnitude of improvement depending on both the degree of orientation jitter and the background color. Random color in the background was found to modulate performance only when color-based processing operated in isolation (jitter 180°). When both contour- and color-processing operated in conjunction (jitter 20°), performance was equally good in the white-background and the random-background conditions. So, an interaction effect was found: the type of background (white versus random color) had a strong effect for jitter 180°, but not for jitter 20°.

There were individual differences in the way subjects utilized the contour smoothness and color cues. Specifically, Subject S2 relied more on color cue so that his performance with only color-based processing (jitter 180° with both white- and random-backgrounds) was higher than when contour-based processing operated in isolation (jitter 20° and no-color condition). In contrast, Subject S1 relied more on contour smoothness cue so that her performance was higher when contour-based processing operated in isolation than when color-based processing operated in isolation. Subjects S3 and S4 fell between the two extremes: color cue alone led to higher performance than contour alone, only with white background.

Nonetheless, when jitter was 20°, all subjects were able to combine the contour and color information to produce similar level of performance. Individual differences were the smallest when both contour and color information were made available.

3 Control experiments

3.1 Effect of distortion

Two subjects, S1 and S2, who performed the experiment using distortion $k = 0.04$ repeated the experiment with $k = 0.08$. Subject S2 was not tested in jitter 180° no-color condition. The results are shown in Figure 5. Increasing egg distortion made the task easier roughly by a factor of two, but the pattern of results is the same as with $k = 0.04$.

The consistent pattern of results for the two distortion coefficients indicated that the same underlying contour integration mechanism was at play for both distortion levels. Therefore, increasing distortion only improves shape discriminability, making the interpretation of the boundaries easier without altering the processing involved in boundary extraction.

3.2 Fixation outside the egg

It was shown that the human visual system could perform local processing based on smooth contours (Field et al., 1993) or color similarity (Kovács, 1996). To determine the role of local versus global processing, we performed a control experiment where fixation was placed outside of the egg. Peripheral viewing precludes the extraction of closed contours using a log-polar representation because the problem can no longer be translated into a shortest path global optimization problem (see Section 4 Model). Therefore,

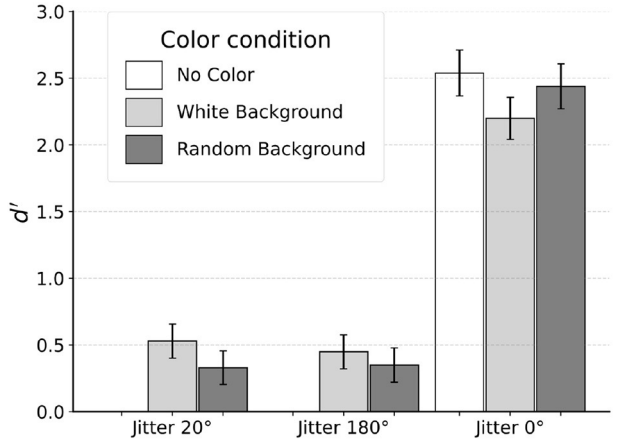


FIGURE 7
Results from peripheral viewing.

we expected local processing to be a critical mechanism during peripheral viewing.

3.2.1 Methods

The same stimulus generation procedure was adopted with the exception that now the center of the egg was randomly placed outside a circle covering the central 50% of the stimulus canvas. This way the fixation was always outside the egg boundary. All six experimental conditions were tested. In addition, three experimental sessions with perfectly smooth contour (jitter 0°) were added. Figure 6 illustrates different egg positions using a target egg with jitter 0°. Subject S1 was tested in this experiment.

3.2.2 Results and discussion

Subject S1 was unable to see the egg in two of the conditions with jitter 20° and jitter 180° that had no color. Thus, no data was actually collected for these conditions. Results from the remaining seven conditions are shown in Figure 7.

For jitter 20°, d' was 0.53 and 0.33 for white-background and random-background conditions respectively. Recall that this subject produced, in the corresponding conditions, d' values of 2.00 and 2.06 when tested with foveal viewing (fixating inside the egg). For jitter 180°, d' was 0.45 and 0.35 for the two color conditions respectively, compared with the d' values of 1.12 and 0.61 with foveal viewing. When tested without orientation jitter (jitter 0°), Subject S1 produced d' values of 2.54, 2.20 and 2.44 for the no-color, white-background and random-background conditions. For comparison, performance was perfect (proportion correct 100%) when jitter 0° was used with foveal viewing.

A decrease in discrimination of checkerboard patterns during peripheral viewing was documented in Schlingensiepen et al. (1986). These authors measured a drop in d' by a factor of two when fixation was outside the stimuli compared to free viewing of the stimuli. In our experiment, fixating outside of the target shape dramatically changed the subject's performance. When tested with jitter 20° and no-color condition, moving the stimulus to the periphery made the stimulus invisible. Adding color helped, only to a small extent. Reliable performance was observed only when smooth contour with jitter 0° was used. We will suggest later in this paper that this change is related to unavailability of the global shortest path optimization when log-polar representation is used. We want to point out that the fact that the egg was invisible in peripheral viewing when there was no color for both jitter 20° and 180° cannot simply be explained by poor visual resolution because the same target shape was clearly visible with jitter 0°.

On top of a decrease in performance, the general pattern of results was different from that with foveal viewing. Specifically, with peripheral viewing, we did not observe an interaction effect between jitter level and the type of background color. These results suggest that the subject had to rely on a completely different mechanism when the fixation was outside the target shape. Specifically, local cues such as smooth contour in the jitter 0° condition had to be used to perform the task. With jitter 20° and greater, local processing based on smoothness is no longer effective. Interestingly, once boundary was smooth, adding color did not improve performance.

We conclude that local processing using contour information could occur in periphery only when sufficiently smooth contours were present. Robustness to orientation jitter could be achieved only when fixation was inside the boundary. Similarly, integration between contour- and color-based processing seemed to occur only when fixation was inside the boundary.

4 Model

4.1 Model architecture

We extended the biologically-inspired model introduced by Kwon et al. (2016) to include color processing. Similar to the model described by Kwon et al. (2016), our model uses the log-polar representation of the image and solves the least-cost path problem by applying Dijkstra algorithm. The log-polar representation is a good approximation to the retinotopic mapping in the early visual areas of the cortex (Schwartz, 1977). By adopting log-polar representation, the computationally hard problem of extracting a closed boundary is transformed into an easier problem of finding the shortest path. Figure 8 provides a graphical summary of the model implementation. Operations that are different compared to the previous model are labeled with asterisks. We describe the individual steps in the following subsections.

4.1.1 Log-polar

The log-polar representation is a good approximation of the mapping from the retina to the visual cortex (Schwartz, 1977). This is because the linear density of ganglion cells in retina is non-uniform, with hyperbolic decrease (approximately $\frac{1}{r}$) as eccentricity, r, increases. Therefore, right at the first step of visual processing is a space-variant sampling of the visual input. This nonuniform sampling is subsequently mapped onto uniformly distributed visual cortical neurons, resulting in an over-representation at the fovea and under-representation at the periphery (cortical magnification). Based

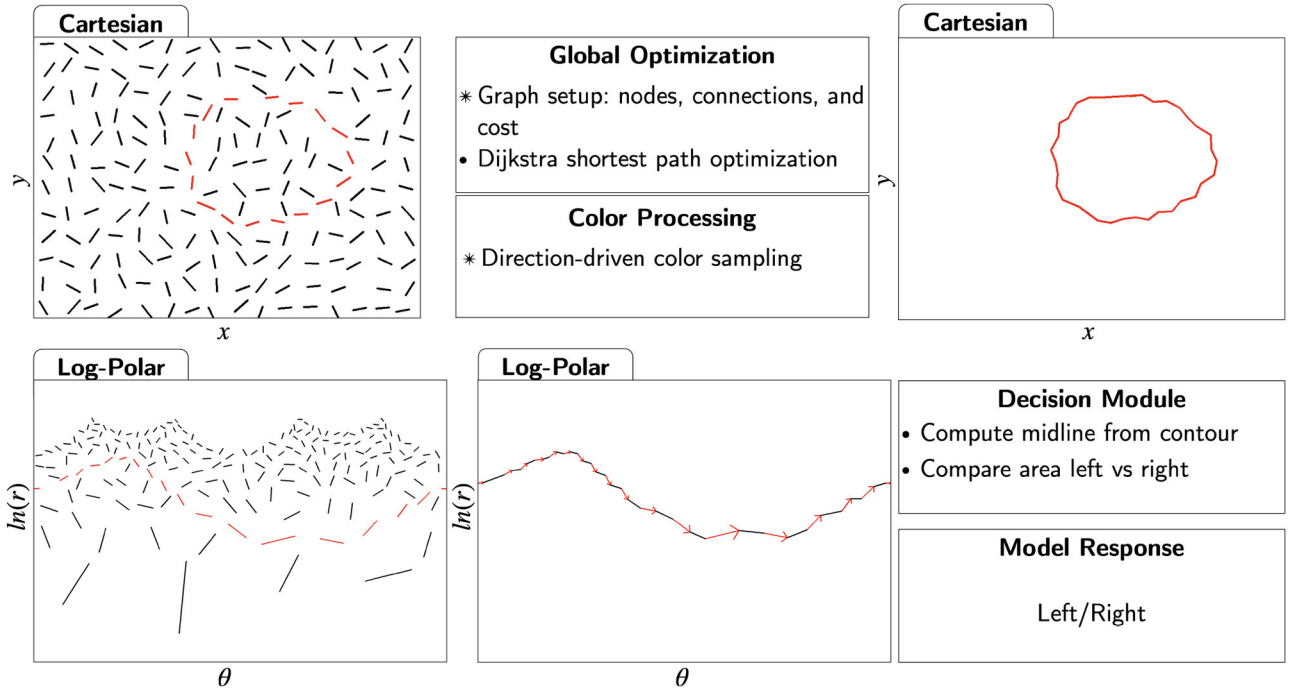


FIGURE 8
Graphical summary of the model. Operations that were modified in this study compared to the previous study by Kwon et al. (2016) are indicated by asterisks. Edges belonging to the egg boundary are highlighted in red. Given an image, the model receives as input a set of edges and the color values associated with each edge. Then, the model fixates at the center of the stimulus canvas and selects an orientation as its reference angle (0 rad.) to define the log-polar coordinate space. A graph is constructed where a node represents the log-polar coordinates of the two endpoints of an edge along with their associated colors on both sides. Nodes are connected to all other nodes in its neighborhood. Global optimization is then performed using Dijkstra shortest path algorithm to connect the representation of the initial node at reference angle (0 rad.) to its representation at 2π rad. The output in Cartesian representation in the form of a closed boundary is obtained by mapping the shortest path solution from log-polar into Cartesian representation. A decision module interprets the boundary output to provide a response if the pointy side of the egg is to the left or to the right.

on measurements of cortical magnification factors in macaques, Schwartz (1977) demonstrated that the log-polar transformation closely approximates this mapping from the retina to the visual cortex.

The log-polar transformation begins with specifying a polar coordinate system on the image. Instead of using Cartesian coordinates (x, y), we use polar coordinates: radius r and angle θ. The origin of the r dimension represents the center of the retina, which is a projection of the point in the visual field where the eye fixates.

The log-polar coordinate system is defined by taking the logarithm of the r dimension. Two requirements must be met for the mapping from Cartesian to log-polar to be the proper transformation as defined in complex analysis: the base of the logarithm must be e (i.e., natural logarithm), and the angle θ must be expressed in radians (not degrees). This way, the log-polar mapping is a conformal mapping, preserving local angles. It is precisely this mapping that has been shown to approximate the mapping from the retina to the primary visual cortex. The best way to avoid confusion is to apply a complex-logarithmic function to the complex variable $(z = x+iy)$ representing an image point (x, y). Any complex number, z, may be expressed in polar form, by using Euler formula: $z = x + iy = r(\cos\theta + i\sin\theta) = re^{i\theta}$. Taking the complex-logarithm of the complex number z, using a logarithm to the base of e results in $log_e(z) = log_e(r) + i\theta$. We would like to point out that software packages or libraries often have a function for log-polar transformation, but this function is not necessarily a conformal mapping, namely the logarithm is not natural and/or angle is not expressed in radians.

4.1.1.1 Cartesian to log-polar transformation

The input stimulus with size [1920 × 1080 pixels] was transformed into a log-polar image with size [1,920 × 1,920 pixels]. The model took as input a set of Cartesian coordinates defining the edges detected in the image. In this paper, we used synthetic images described in Section 2.1.1. As a result, our model did not have to perform edge detection because the edges already existed. The model took as input an $[N \times 4]$ matrix where N is the number of edges in the stimulus, and each edge was defined by its two endpoints in the Cartesian coordinates, (x_1, y_1, x_2, y_2).

We then transformed the Cartesian coordinates into log-polar coordinates. We defined $r = 0$ to be the fixation cross, which was placed at the center of the stimulus image. The origin for the polar angle was selected using the same strategy as in the previous study (Kwon et al., 2016). Specifically, $\theta = 0$ was set at the midpoint of a randomly selected starting edge belonging to the boundary of the target egg. This edge was used as the start/end point for computing the shortest path. Kwon et al. (2016) showed that if a starting point was not provided, the model could try a number of starting points and compute the shortest path for all these points. The shortest

path from all these paths almost always corresponded to the correct boundary (see their Model LI-SP-EST).

4.1.2 Global optimization

With the new representation in the log-polar space, the original task of boundary extraction was framed as finding the shortest path connecting the representation of the starting edge at 0 rad. back to its representation at 2π rad. Dijkstra shortest path algorithm was used to perform global optimization. Below, we describe the three main components in setting up a graph for optimization: defining the nodes in the graph, establishing connections between nodes, and assigning the costs of travel from one node to another. Connection between nodes is more commonly termed as an "edge" in the context of graph theory. To avoid conflict in terminology, we use the term "edge" when discussing a detected edge in the image; we use the term "connection" to mean the edge from one node to another in a graph.

4.1.2.1 Defining a node in the graph

We defined a node in the graph using three sets of values: an ordered set of endpoints of an edge in the log-polar coordinates $[(r_1, \theta_1), (r_2, \theta_2)]$, and two sets of color values associated with the region to the left and to the right of the edge, RGB (Left), RGB (Right).

To encode contour-related information in the graph, the position and orientation of a detected edge was included in the definition of a node as an ordered set of endpoints in the log-polar coordinates $[(r_1, \theta_1), (r_2, \theta_2)]$. Since there are two possible directions of travel between two endpoints, each log-polar edge was represented twice in the graph: once for the forward direction, $[(r_1, \theta_1), (r_2, \theta_2)]$, and the other for the reversed direction, $[(r_2, \theta_2), (r_1, \theta_1)]$. A similar implementation where an edge was represented twice in order to explicitly express direction was described by Williams and Thornber (1999).

Next, we describe our approach to introduce color-related information in the graph. In particular, we would like to encode color in a way that would preserve the contour-color relationship (Rentzeperis et al., 2014). As an illustration of the contour-color relationship, imagine a white circle placed on a black background. The closed boundary of the circle separates the stimulus canvas into two regions, foreground and background. The region with white color coincides with the area enclosed by the boundary. Therefore, color information does not contradict the contour-defined boundary. In order to distinguish foreground color from background color while respecting contour edges, we propose the notion of directionality, being inspired by Stahl and Wang (2007). Imagine walking on the boundary of a circle clockwise. The white color belonging to the interior region of the circle is always to the right at the walker local frame. Considering the direction of travel allows the two pieces of information from contour and color to be tracked simultaneously: for contour, the orientation of an edge is the unsigned direction; for color, color similarity in the foreground versus background can be tracked by comparing the colors on both sides of an interpolating edge. In Figure 9, the shaded regions indicate the regions to the left of edges according to their respective directions of travel.

For each edge, color was sampled from a Moore neighborhood of range three (7×7 grids) in the Cartesian representation (Moore, 1964). Colors to the left and right of the edge were averaged separately to obtain two sets of RGB values.

4.1.2.2 Defining connections in the graph

We restricted the connectivity in the graph, so that a node in the graph can reach only the set of nodes located within its neighborhood. We defined a neighborhood as a square window of 240×240 pixels in the log-polar representation. Therefore, instead of constructing a complete graph with connections for every pair of nodes, only nodes that were sufficiently close to each other were connected. Our pilot tests showed that the quality of solutions was not affected, but computation time was greatly improved.

4.1.2.3 Cost of interpolation

To calculate the cost of interpolation (the cost of a connection in the graph) from Node A to Node B, we used the following features: (1) distance, (2) turning angle, (3) color similarity, and (4) color contrast. The value of each feature was multiplied by its weight. We describe the algorithmic computation for each feature, as well as their relationships to the computational level representation of Gestalt principles (Marr, 2010).

Let us begin with contour information encoded in the edges: distance and turning angle. The visual system is more likely to choose a particular interpolation if the distance (length of the interpolation) is short, commonly referred to as the Gestalt principle of proximity (Wertheimer, 1923). We computed the distance as the Euclidean distance from the second endpoint of the first node to the first endpoint of the second node. Since distance is computed after the log-polar transformation, scaling a shape has no effect on the distance metric. This behavior is desirable since proximity principle has been shown to be robust to transformations of scaling (Kubovy et al., 1998). We squared the interpolated distance in the cost function to progressively penalize long interpolations. This produced good results, but the actual shape of this function (polynomial vs. exponential) should be tested in the future.

Turning angle was used in the cost of interpolation because smaller changes in orientation are more likely to be interpolated (Wertheimer, 1923; Elder, 2018). Turning angle was defined as $\psi = |\psi_1| + |\psi_2|$, where ψ_1 is the angle formed by the first endpoint of Node A, second endpoint of Node A, and first endpoint of Node B; and ψ_2 is the angle formed by the second endpoint of Node A, first endpoint of Node B, and second endpoint of Node B. The angles ψ_1, ψ_2 are also labeled in Figure 9. Minimizing turning angle minimizes abrupt changes in the direction of travel, and thus encodes the Gestalt principle of good continuation.

As a natural consequence of minimizing the turning angle in the log-polar space, Gestalt principle of convexity is encoded implicitly without including additional parameter in the cost function. An easy way to see this is to realize that a circle around the fixation point maps into a straight line in the log-polar representation.

Next, we discuss features related to color-based processing: color similarity and color contrast. Edges are more likely to be connected when they share the same colors on the left side and/or

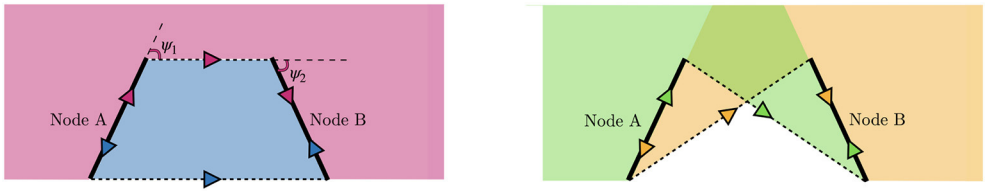


FIGURE 9
There are four possible permutations representing the interpolation between two edges in the image. Solid line denotes a node in a graph with the direction of travel indicated by the arrow. Dashed lines denote the interpolations. The two angles of interpolation, ψ_1 and ψ_2 are marked. Turning angle is the sum of the absolute values of these angles.

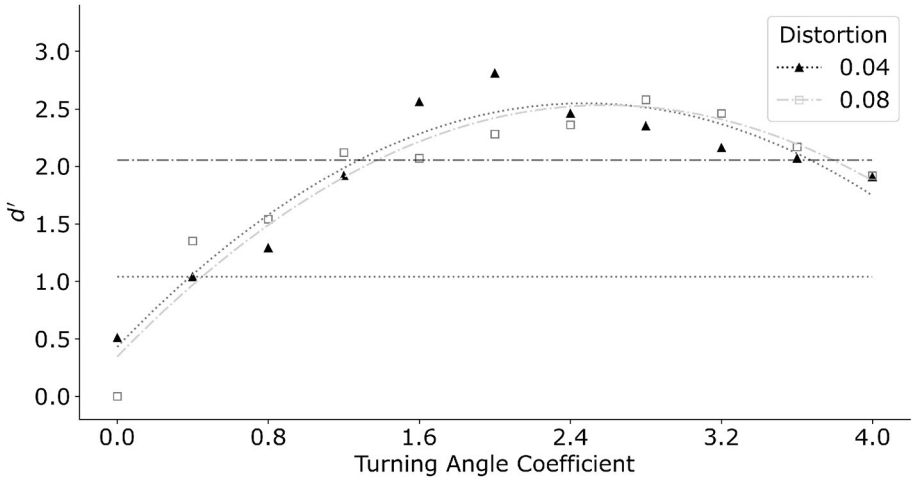


FIGURE 10
Model performance as a function of increasing turning angle coefficient. Model performance improved initially as turning angle coefficient increased. Further increase of turning angle coefficient degraded performance. Horizontal dotted line and dash-dot line indicate performance of an average subject for distortion $k = 0.04$ and $k = 0.08$ respectively. Turning angle coefficient of 0.4 and 1.2 produced performance closest to that of an average subject, and thus was chosen for subsequent simulations for conditions with color.

on the right side of the interpolated curve, also described as the Gestalt principle of similarity (Kovács, 1996). The difference in colors between the left side of Node A and left side of Node B were computed as follows: Δ Color(Left) = $|\text{RGB(Left)}_A - \text{RGB(Left)}_B|$, and similarly for the right sides of both nodes Δ Color(Right) = $|\text{RGB(Right)}_A - \text{RGB(Right)}_B|$. The two differences were combined using a minimum operation, Color similarity = min(Δ Color(Left), Δ Color(Right)). As a result, a pair of edges is considered to share similar color as long as they share similar colors on at least one side.

Finally, edges that carry higher color contrasts between the left and right side are more likely to indicate the presence of a boundary. High contrast relates to the pop-out effect or the Gestalt principle of dissimilarity (Pinna et al., 2022). We compared the colors on both sides of an individual node and computed the color contrast = $|\text{RGB(Left)}_A - \text{RGB(Right)}_A|$. We used the negative of color contrast for global minimization. Note that the notion of directionality was not encoded in the computation of color contrast, since the two nodes representing the same edge in both directions have the same value for color contrast.

The total cost for every connection in the graph was calculated by summing the cost across the four features, with weights defined by coefficients. The cost function with their normalizing constants was as follow:

$$a_1(\text{D}^2/1920) + a_2(\text{TA}/2\pi) + a_3(\text{CS}/255) + a_4(1 - \text{CC}/255)$$

where a_1, a_2, a_3, a_4 are the coefficients of the individual features; D, TA, CS and CC represent the cost of distance, turning angle, color similarity, and color contrast respectively. Since changing the coefficients alters the model behavior, we identify a model by specifying its coefficients. For example, a model ignoring color information would set the coefficients for color similarity and color contrast to zero. For the ease of reporting, we label the model in terms of their coefficients using the following convention [distance, turning angle, color similarity, color contrast]. If the model assigned a coefficient of 1 to both distance and turning angle, it would be labeled as [1,1,0,0].

Note that the magnitudes of the four features in the cost function were very different because of the units used (distance was measured in pixels, angle in radians, and color using 256 digital units). As a result, the values of these features were rescaled by their respective constants to be in comparable ranges. This means that the values of coefficients should not be interpreted literally:

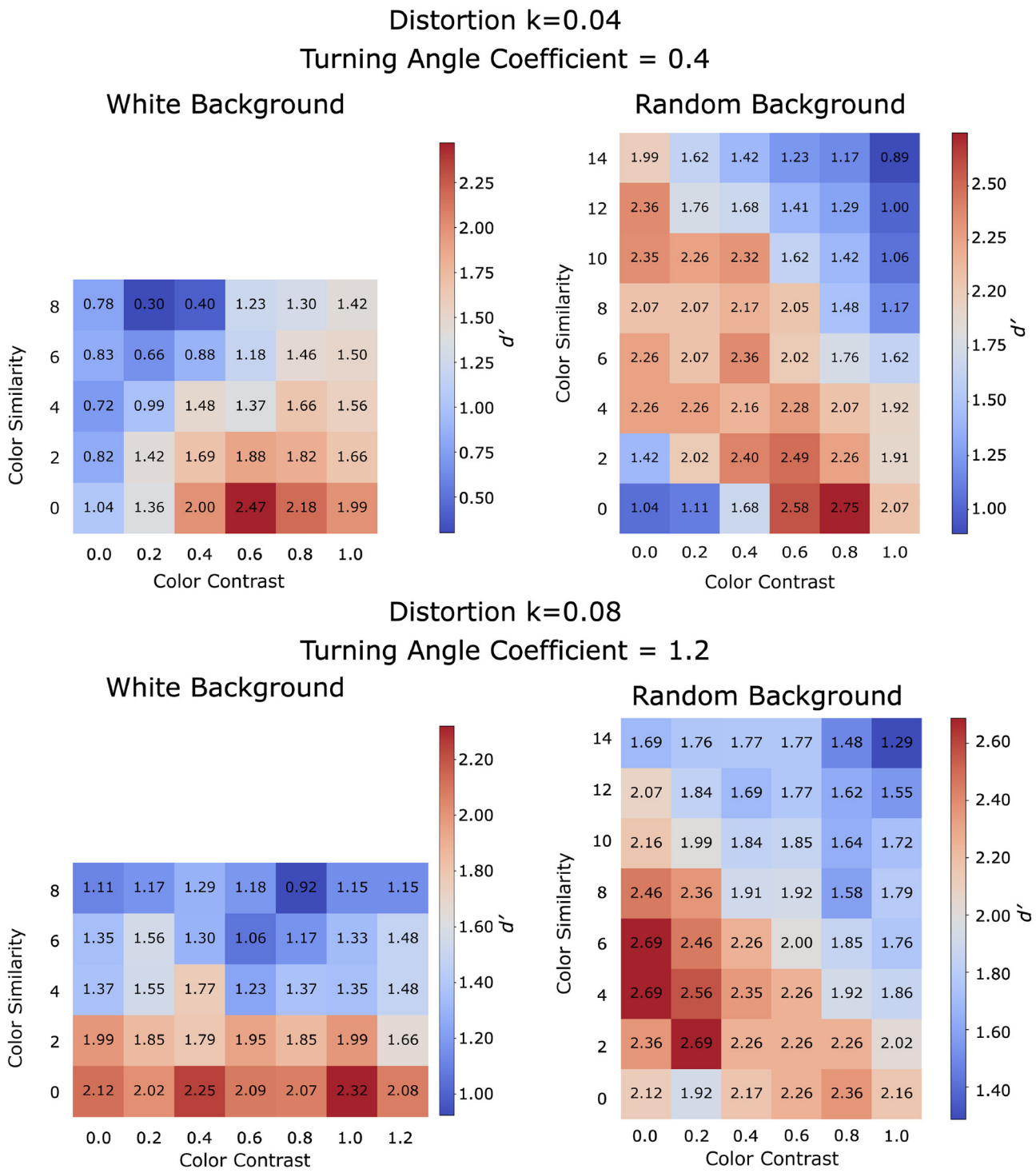


FIGURE 11
The model parameter space by performing grid search on stimuli with jitter 20°. Model performance was evaluated using d'. Results are color coded with red indicating high performance, grey indicating intermediate performance, and blue indicating low performance. **Top**: Grid search results for distortion $k = 0.04$. **Bottom**: Grid search results for distortion $k = 0.08$.

e.g. a coefficient value of one for distance and a coefficient value of ten for color similarity does not mean that color similarity is ten times more important than proximity. It is, however, possible to make relative comparisons of the components of the cost function across different conditions: an increase in coefficient for color similarity from a value of two to a value of four (assuming that other coefficients stayed the same) meant that color similarity was twice as important in the second condition than the first.

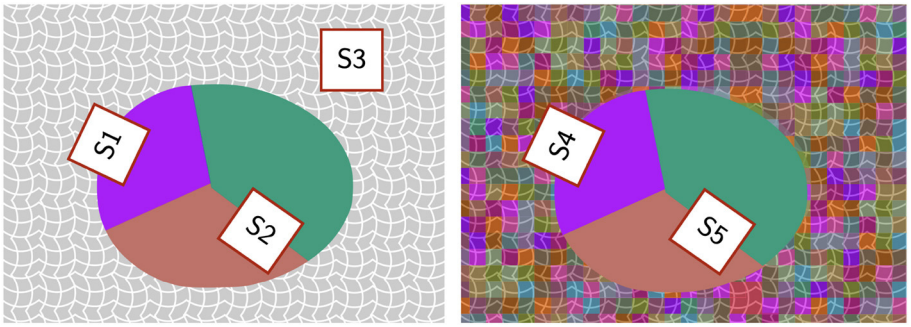
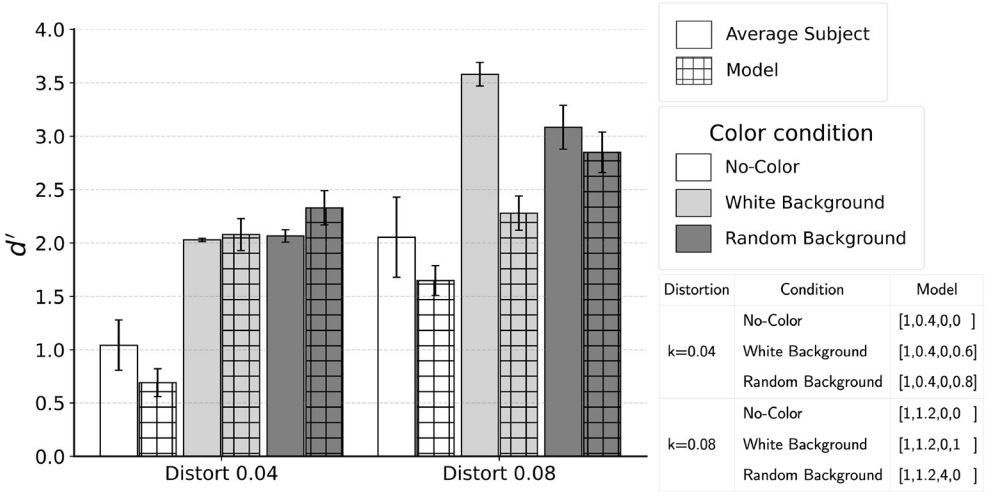


FIGURE 12
Schematic illustration comparing color similarity and color contrast at different regions of a stimulus. **Left**: White background condition. **Right**: Random background condition. Details described in text.



Distortion	Condition	Model
k=0.04	No-Color	[1,0.4,0,0]
	White Background	[1,0.4,0,0.6]
	Random Background	[1,0.4,0,0.8]
k=0.08	No-Color	[1,1.2,0,0]
	White Background	[1,1.2,0,1]
	Random Background	[1,1.2,4,0]

FIGURE 13
Comparison of the model and subject performance for jitter 20°. Model was able to match performance of an average subject except in one condition with higher distortion $k = 0.08$ and white background. Refer to text for discussion on the interpretation and a plausible model which could perform better for this condition.

4.1.3 Producing closed contour as model output

After setting up the graph by defining the nodes, connections, and costs, we applied Dijkstra shortest path algorithm. The algorithm solved a global optimization to produce a least-cost path from a starting node to itself. After the shortest path was transformed into edges in the Cartesian representation, pairs of edges were interpolated using straight line segments to produce a closed boundary. The literature provides more sophisticated interpolation methods that could be used in our model (Sharon et al., 2000; Kimia et al., 2003; Stahl and Wang, 2007; Kalar et al., 2010; Ben-Yosef and Ben-Shahar, 2011; Singh, 2015).

The model guarantees closure because the two endpoints of the least-cost path in log-polar translate to the same point in Cartesian space. Therefore, the boundary extraction solution from the model aligns with the Gestalt principle of closure, such that closed contours are perceptually preferred over open ones (Kovacs and Julesz, 1993).

In summary, by representing the problem in log-polar space and performing global optimization using the proposed cost function, a total of six Gestalt principles were operationalized. They are proximity, good continuation, convexity, color similarity, color dissimilairty, and closure.

4.1.4 Decision module

To produce a response to the 2AFC question, we adopted the same decision criterion used by Kwon et al. (2016). Specifically, the model took the horizontal range of the detected boundary and computed the midpoint. The extracted boundary was divided into two areas by drawing a vertical line. The area of the left half was compared to the area of right half. The pointy side of the egg was the side with a smaller area.

4.2 Selecting the parameters of the model

The behavior of the model is determined by four parameters, namely the set of coefficients weighing the four features in the

cost function (distance, turning angle, color similarity and color contrast). To examine the effect of each model parameter, we created a separate set of 100 egg stimuli and performed grid search on the experimental conditions. Model performance was estimated using d'. Since the main goal of the current study was to explore the integration of contour and region (color), our simulations focused on the conditions with jitter 20°. As explained in Section 2.1.1, jitter 180° completely removed contour information by randomizing the orientations of the edges belonging to the egg boundary. It follows that the subjects had to rely exclusively on the color information. We will analyze this condition in the future.

We first established a baseline for the model performance on the no-color condition using contour-related features (distance and turning angle). We did this for both distortion coefficients: $k = 0.04$ and $k = 0.08$.

Interpolation distance is the dominant feature in the model. The previous version of the log-polar model reported in Kwon et al. (2016) used only the interpolation distance combined with a linear interpolation front-end. Their linear interpolation formed longer contours by connecting approximately collinear edges within a small neighborhood before finding the least-cost path. In the current model, turning angle replaced the linear interpolation. Figure 10 shows the effect of the turning angle coefficient relative to the distance coefficient for two distortions of the egg shape. In this graph we varied the turning angle coefficient from zero to four with a step size of 0.4. Distance coefficient was set to one. Filled triangles represent distortion $k = 0.04$ while open squares represent $k = 0.08$. The same model produced similar performance for both distortion coefficients of the egg. In general, manipulating turning angle coefficient produced a systematic change in performance. The maximum d' was produced at turning angle coefficient of about two. Performance degraded for larger values of the turning angle coefficient. This was related to the model making long interpolations in the log-polar map, producing circular-like parts that did not approximate the egg shape well.

The turning angle coefficients which best captured performance of an average subject was 0.4 for the smaller egg distortion $k = 0.04$, and 1.2 for $k = 0.08$. We therefore fix the turning angle coefficients at the respective values in the subsequent tests which included color. Note that in the main experiment with $k = 0.04$, the three subjects produced d' varying between 0.5 to 1.5. The best performance was produced by S1 who received substantially more practice with these stimuli.

Using the turning angle coefficient identified for each egg distortion (0.4 and 1.2 for distortion $k = 0.04$ and $k = 0.08$ respectively), we performed grid search on color similarity and color contrast coefficients for both white- and random-background conditions (the distance coefficient was set to 1 for all simulations). This grid search was informed by a pilot study exploring a wider range of coefficients. The results for distortion $k = 0.04$ and distortion $k = 0.08$ are summarized in Figure 11. Manipulating color-related coefficients resulted in a gradual change in performance, indicating that the model is stable. For both color conditions, the model was able to combine at least one color feature with contour information to arrive at a higher performance than in the no-color condition. Model performance in the no-color condition is represented by the grid cell where both color similarity and color contrast coefficients are set to zero (bottom left corner of each grid).

We will describe the role of color-related coefficients for each color condition separately. For the white-background condition, the model performed well by using positive coefficients for color contrast while ignoring color similarity: increasing color similarity coefficient degraded performance. Since the stimuli consisted of a Worley-colored egg embedded in white background, information about the target shape can be captured well by the color contrast between the inside and outside of the egg (region S1 in Figure 12). Although color contrast could also be high at region S2, the Gestalt principles of convexity and good continuation will bias the solution towards the egg boundary. Increasing the color similarity coefficient (i.e., penalizing color dissimilarity on each side of the contour) also increases the preference to produce a contour passing through the uniformly white noise edges in the background (region S3 in Figure 12). It is important to point out that our white-background condition is computationally simple (see Figure 12, second row), because the model could remove (filter out) all white edges and would be able to extract the shape boundary nearly perfectly with performance close to perfect (the actual performance will not be perfect because the edges of the egg boundary had 20° random orientation jitter). We verified this directly, but the grid search was done without removing white edges in the background.

For the random-background condition, the model can produce high performance for a range of color coefficients. Specifically, increasing color similarity coefficient or increasing color contrast coefficient could both improve performance (Figure 11). This is illustrated by region S4 in Figure 12 where there is color contrast across the boundary and color similarity for the region inside the egg. However, our exploration showed that color-related parameters are limited in their utility. For example, both color similarity and color contrast parameters could bias the solution towards the polygonal shapes inside the egg with Worley color pattern (region S5 in Figure 12) because the polygonal boundaries have high color similarity on both sides and high color contrast across. This could lead to errors in contour integration. Further research is needed to investigate the role of color-related processing in boundary extraction, especially when color introduces geometrical patterns conflicting with the target boundary (e.g., in the case of camouflage).

The results from the parameter space explorations (Figure 11) suggest that the model was able to integrate color with contour features to arrive at a higher performance when compared to contour alone. This improvement was higher with the more difficult case where egg distortion $k = 0.04$. The model coefficients that led to good performance in the grid search were tested in the next section, using the same stimuli which subjects were tested on.

4.3 Comparing the model to psychophysical results

Based on the results presented in Figure 11, we applied the model to the images that were shown to subjects for the conditions with jitter 20° with both distortion $k = 0.04$ and 0.08. Because the grid search based on 100 images showed that high performance was

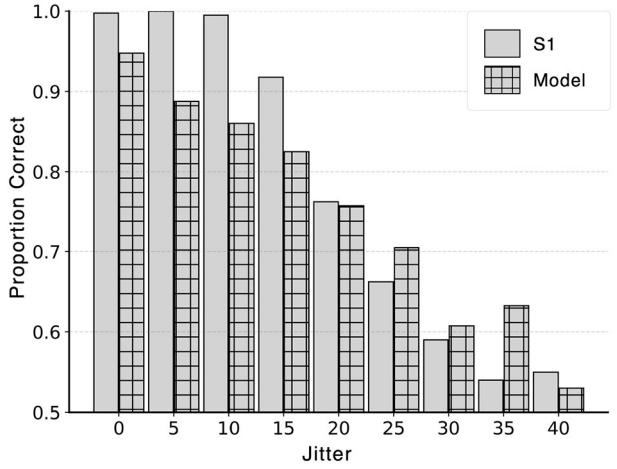


FIGURE 14
Control experiment investigating the effect of jitter on performance of model and subject. Performance decreased from close to perfect (1.0) to chance (0.5) for both model and subject. Standard error for each bar was no greater than 0.025.

achieved with several sets of the coefficients, we applied these sets of coefficients to the 400 images from psychophysical experiment. Differences in model performance across these sets of coefficients were small.

In Figure 13, we report the performance for the model coefficients which produced the highest performance in the grid search. The model was successful in matching human performance for five of the six conditions. For distortion $k = 0.08$ white-background condition, subjects performed close to perfect while the model did not. It is possible that the uniform white background noise edges allowed for simple filtering operations to remove the background before extracting contour. We tested the possibility of such pre-processing by applying the model after all white background edges were removed. Using coefficients [1,1.2,0,0], the model produced $d' = 3.56$, which is almost identical to the average performance of the two subjects who were tested in the control experiment. Future studies can test the hypothesis that the visual system applies a filtering front-end.

To summarize, our results showed that the log-polar based model was successful in integrating contour and color in the test with the egg-like stimuli. The model's performance was not very different from the subjects'. We want to point out that the task was computationally difficult for several reasons: (i) the contour of the egg was fragmented to have a support ratio of 0.5; (ii) the density of the background edges was the same as the density of the edges representing the egg; (iii) the edges representing the contour had random jitter which essentially excluded spatially local growth of the contour based on smoothness; (iv) global optimization was necessary while at the same time avoiding combinatorial explosion related to examining all subsets of edges in the image; (v) the contour had to be closed. Therefore, it is probably not surprising that the model's performance did not exceed that of the best subject. We are confident that our model captured something important about the visual mechanisms of contour integration. However, several components of the model could be further developed and produce even better fit to the subjects' results. To further examine the correlation between the model and the subject, we performed an additional control experiment manipulating jitter level (Section 4.3.1).

4.3.1 Control experiment: effect of jitter

In this control experiment, jitter level was manipulated from 0° to 40° with a step size of 5°. It was natural to expect that increasing jitter (producing non-smooth contours) will lower the performance of subjects. A model which explains how the visual system works would also be affected by jitter level, with a similar degree of quantitative effect.

The stimuli for the control experiment were generated using the procedure for distortion $k = 0.04$, No-Color condition described in Section 2.1.1. All target edges in a particular jitter level had a random change in orientation in the range [-jitter - 5°, -jitter + 5°] or [+jitter - 5°, +jitter + 5°] except for jitter level 0°, where no random jitter was added to the target edges. Subject S1 ran additional eight sessions excluding jitter 20°, which was performed as part of the main experiment (Section 2.2). The model with coefficients 1 for distance and 0.8 for turning angle was chosen because the model produced similar performance (d') as subject S1 based on the simulation in Figure 10. The model was tested on all nine conditions and a comparison between model and subject performance is shown in Figure 14.

Performance of both the model and S1 decreased with increasing jitter level. Pearson correlation coefficient between the model and S1's proportion correct was high: R = 0.96 (p-value<1E-4). Figure 14 shows overall proportion correct instead of d' because d' approached infinity when S1's performance was close to perfect (either because there was no miss or false alarm) for the first three jitter levels below 15°[3]. This high performance could partially be attributed to the role of local interpolation in boundary extraction. The current model relies exclusively on spatially global optimization. Therefore, adding local interpolation as the front-end would likely produce close to perfect model performance for small jitter levels (see Figure 1). A drop in S1's performance was observed with jitter 15° and above, suggesting that local operations failed with high jitter levels. This result validated the choice of 20° jitter to investigate the role of global processing.

5 Conclusion

Given a 2D retinal or camera image, determining which contour and region belong to a single object is the first step to recognizing the object and reconstructing its 3D shape. Our psychophysical experiments eliminated local contour cues by introducing orientation jitter to explore the interaction between edge-based and color-based processing in the context of global processing. We showed that each of these two types of processing could operate in isolation: edge-based processing could reliably extract boundaries when contours were relatively smooth; and color-based processing could reliably extract boundaries when

3 Nevertheless, d' follows a similar pattern: higher jitter leads to systematically lower d'.

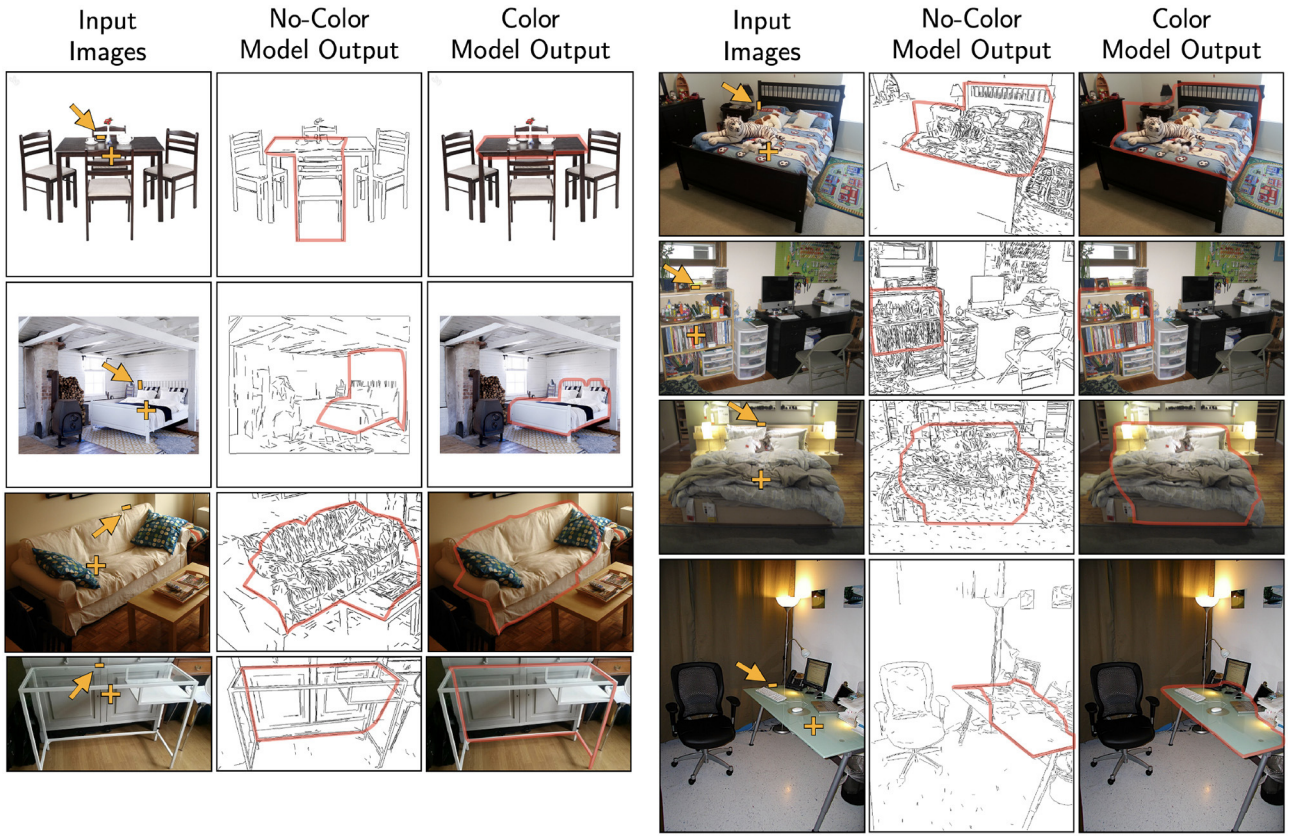


FIGURE 15
Model performance tested using real images of furniture. The fixation point and starting edge were marked. **Left**: shows the cases where the model produced a different output when color information was used. **Right**: illustrates the stability of model such that adding color produced minimal differences when the no-color model could produce reasonable outputs.

color in the foreground was different than the color in the background. When both contour and color cues were present, subjects were able to integrate the two pieces of information to produce the highest performance. We established these results under viewing conditions where the subject fixated inside the boundary of the object. Moving the fixation outside the boundary substantially impaired subject's performance, but this impairment cannot be explained by a lower visual resolution in the periphery. The design of our experimental stimuli may be extended in future studies. One may manipulate (i) the target 2D shape, (ii) support ratio of the fragmented contour, (iii) the degree of similarity between color inside and outside of the 2D shape, and (iv) the texture pattern for the target shape and the background. These characteristics represent conventional features that have been used to study figure-ground organization.

We proposed a biologically-inspired boundary extraction model combining contour-based processing with color-based processing. The model was tested on the conditions with 20° jitter and its performance was similar to that of the subjects. The main characteristic of the model is the use of the log-polar representation which is known to be a good approximation of the retinotopic mapping in the primary visual areas of the brain. By performing shortest path optimization in the log-polar representation, the model performed global optimization to produce a boundary solution which is guaranteed to close. The model integrated two contour-related features (distance of interpolation and turning angle) and two color-related features (color similarity and color contrast) in its cost function. More specifically, the interaction between contour and color was modeled using the concept of boundary directionality, where the model encoded color as guided by contour-based cues. The model produced reliable results comparable to that of subjects with the two difficult conditions of no color and random background when jitter was 20°. We hope that these results will stimulate further explorations of competing boundary extraction models.

In order to reveal the relationship between contour and color, our present study used synthetic images to increase the difficulty of the task so that subjects would not perform at ceiling. Using synthetic stimuli also allowed us to manipulate contour and color cues independently to control the difficulty across conditions. Since the model could replicate human contour-color interaction using these difficult synthetic stimuli, one could expect that the model should be able to extract boundaries using real world images, which typically are easy for human observers. Without any additional tuning of the model parameters, two versions of the model (no-color and color) were applied to real images of furniture from the Pix3D dataset (Sun et al., 2018). Specifically, we used the

coefficients from the jitter 20° random-background condition, which better resembled the amount of noise in real images. Two sets of coefficients were tried: [1,0.4,0,0.8] from distortion $k = 0.04$ and [1,1.2,4,0] from distortion $k = 0.08$. Both sets of coefficients produced similar outputs. Given an input image, an additional pre-processing stage of edge detection was applied (Lee et al., 2014). The no-color model received only the detected edges as input; whereas the color model received additional color information associated to the left and right regions of each edge. Fixation point was placed inside the shape, and a random edge belonging to the target shape was given as the initial edge. Figure 15 shows examples of the model output tested on five different categories of furniture: table, bed, sofa, desk, and bookcase. The preliminary results suggest that the model could be applied to a wide variety of real images. Future studies could test the model generalizability by using real images from different domains.

Another topic for future research is to integrate saliency maps with the model. Because the model used the log-polar representation, there is a requirement for the fixation point to be placed inside the boundary of the target object. Previous literature has suggested that humans use sophisticated attention mechanisms to guide fixation, Schütz et al. (2011), one example being the salience network for bottom-up processing. This network integrates different features such as orientation, color, or motion to create a saliency map which highlights the regions in the image that are most relevant for fixation (for a review, see Uddin, 2016).

Finally, the boundary extraction model could be used as a front-end model for higher order visual processing such as 3-dimensional (3D) object reconstruction. We already showed that if the symmetry correspondence problem is solved, 3D shape recovery can be accomplished (Pizlo et al., 2014). However, solving 3D symmetry correspondence for several objects in a 2D camera image is computationally challenging, if possible at all. Restricting the symmetry correspondence analysis to one object at a time will be likely to produce acceptable solutions.

Data availability statement

Stimuli used for this study can be found on The Open Science Framework: osf.io/fq5hu, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by UCI Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

DH and ZP contributed to conception and design of the study and contributed to manuscript preparation. DH collected and analyzed psychophysical data, formulated the model, and performed simulations. All authors contributed to the article and approved the submitted version.

Acknowledgments

We thank Dr. Jordan Rashid for measuring the monitor color gamut. We thank Mark Beers for his comments that improved our stimuli. We also thank the editor and the reviewers whose comments and suggestions allowed us to improve the paper.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Bednar, J. A., Signell, J., Kbowen, Chris, B., Sutherland, D., Stevens, J.-L., and Kats, P. (2020). *Holoviz/Colorcet: Version 2.0.2*. Zenodo. doi: 10.5281/zenodo.3929798

Ben-Yosef, G., and Ben-Shahar, O. (2011). A tangent bundle theory for visual curve completion. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 1263–1280. doi: 10.1109/TPAMI.2011.262

Elder, J. H. (2018). Shape from contour: Computation and representation. *Ann. Rev. Vision Sci.* 4, 423–450. doi: 10.1146/annurev-vision-091517-034110

Field, D. J., Hayes, A., and Hess, R. F. (1993). Contour integration by the human visual system: evidence for a local "association field". *Vision Res.* 33, 173–193. doi: 10.1016/0042-6989(93)90156-Q

Grossberg, S., and Mingolla, E. (1985). Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading. *Psychol. Rev.* 92, 173. doi: 10.1037/0033-295X.92.2.173

Hansen, T., and Gegenfurtner, K. R. (2009). Independence of color and luminance edges in natural scenes. *Vis. Neurosci.* 26, 35–49. doi: 10.1017/S0952523808080796

Kalar, D. J., Garrigan, P., Wickens, T. D., Hilger, J. D., and Kellman, P. J. (2010). A unified model of illusory and occluded contour interpolation. *Vision Res.* 50, 284–299. doi: 10.1016/j.visres.2009.10.011

Kimia, B. B., Frankel, I., and Popescu, A.-M. (2003). Euler spiral for shape completion. *Int. J. Comput. Vis.* 54, 159–182.

Kovács, I. (1996). Gestalten of today: Early processing of visual contours and surfaces. *Behav. Brain Res.* 82, 1–11. doi: 10.1016/S0166-4328(97)81103-5

Kovacs, I., and Julesz, B. (1993). A closed curve is much more than an incomplete one: effect of closure in figure-ground segmentation. *Proc. Nat. Acad. Sci.* 90, 7495–7497. doi: 10.1073/pnas.90.16.7495

Kozma-Wiebe, P., Silverstein, S. M., Fehér, A., Kovács, I., Ulhaas, P., and Wilkniss, S. M. (2006). Development of a world-wide web based contour integration test. *Comput. Human Behav.* 22, 971–980. doi: 10.1016/j.chb.2004.03.017

Kubovy, M., Holcombe, A. O., and Wagemans, J. (1998). On the lawfulness of grouping by proximity. *Cogn. Psychol.* 35, 71–98. doi: 10.1006/cogp.1997.0673

Kwon, T., Agrawal, K., Li, Y., and Pizlo, Z. (2016). Spatially-global integration of closed, fragmented contours by finding the shortest-path in a log-polar representation. *Vision Res.* 126, 143–163. doi: 10.1016/j.visres.2015.06.007

Lee, J. H., Lee, S., Zhang, G., Lim, J., Chung, W. K., and Suh, I. H. (2014). "Outdoor place recognition in urban environments using straight lines," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. Hong Kong: IEEE, 5550–5557.

Macmillan, N. A., and Creelman, C. D. (2004). *Detection Theory: A User's Guide*. London: Psychology Press. doi: 10.4324/9781410611147

Marr, D. (2010). *Vision: A Computational Investigation into the Human Representation And Processing of Visual Information*. Cambridge, MA: MIT Press.

Moore, E. F. (1964). *Sequential Machines: Selected Papers*. Reading, Massachusetts: Addison-Wesley Longman Ltd.

Moutoussis, K. (2015). The physiology and psychophysics of the color-form relationship: a review. *Front. Psychol.* 6, 1407. doi: 10.3389/fpsyg.2015.01407

Pinna, B., Brelstaff, G., and Spillmann, L. (2001). Surface color from boundaries: a new 'watercolor' illusion. *Vision Res.* 41, 2669–2676. doi: 10.1016/S0042-6989(01)00105-5

Pinna, B., Porcheddu, D., and Skilters, J. (2022). Similarity and dissimilarity in perceptual organization: On the complexity of the gestalt principle of similarity. *Vision* 6, 39. doi: 10.3390/vision6030039

Pizlo, Z., Li, Y., and Sawada, T. (2014). *Making a Machine That Sees Like Us*. Oxford: Oxford University Press, USA.

Rentzeperis, I., Nikolaev, A. R., Kiper, D. C., and van Leeuwen, C. (2014). Distributed processing of color and form in the visual cortex. *Front. Psychol.* 5, 932. doi: 10.3389/fpsyg.2014.00932

Rosenfeld, A., and Thurston, M. (1971). Edge and curve detection for visual scene analysis. *IEEE Trans. Comp.* 100, 562–569. doi: 10.1109/T-C.1971.223290

Schlingensiepen, K.-H., Campbell, F., Legge, G. E., and Walker, T. (1986). The importance of eye movements in the analysis of simple patterns. *Vision Res.* 26, 1111–1117. doi: 10.1016/0042-6989(86)90045-3

Schütz, A. C., Braun, D. I., and Gegenfurtner, K. R. (2011). Eye movements and perception: A selective review. *J. Vis.* 11, 9–9. doi: 10.1167/11.5.9

Schwartz, E. L. (1977). Spatial mapping in the primate sensory projection: analytic structure and relevance to perception. *Biol. Cybern.* 25, 181–194. doi: 10.1007/BF01885636

Sharon, E., Brandt, A., and Basri, R. (2000). Completion energies and scale. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 1117–1131. doi: 10.1109/34.879792

Singh, M. (2015). "Visual representation of contour and shape," in *Oxford Handbook of Perceptual Organization* (Oxford Academic), 236–258.

Stahl, J. S., and Wang, S. (2007). Edge grouping combining boundary and region information. *IEEE Trans. Image Proc.* 16, 2590–2606. doi: 10.1109/TIP.2007.904463

Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., et al. (2018). "Pix3D: Dataset and methods for single-image 3D shape modeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2974–2983.

Tanimoto, S., and Pavlidis, T. (1975). A hierarchical data structure for picture processing. *Comp. Graph. Image Proc.* 4, 104–119. doi: 10.1016/S0146-664X(75)80003-7

Taylor, J., and Xu, Y. (2022). Representation of color, form, and their conjunction across the human ventral visual pathway. *Neuroimage* 251, 118941. doi: 10.1016/j.neuroimage.2022.118941

Taylor, J., and Xu, Y. (2023). Comparing the dominance of color and form information across the human ventral visual pathway and convolutional neural networks. *J. Cogn. Neurosci.* 35, 816–840. doi: 10.1162/jocn_a_01979

Tyler, C. W., and Solomon, J. A. (2019). Color perception in natural images. *Curr. Opin. Behav. Sci.* 30, 8–14. doi: 10.1016/j.cobeha.2019.04.002

Uddin, L. Q. (2016). *Salience Network of the Human Brain*. Cambridge, MA: Academic Press.

Vergeer, M., Anstis, S., and van Lier, R. (2015). Flexible color perception depending on the shape and positioning of achromatic contours. *Front. Psychol.* 6, 620. doi: 10.3389/fpsyg.2015.00620

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt II. *Psychol. Forsch.* 4, 301–350. doi: 10.1007/BF00410640

Williams, L. R., and Thornber, K. K. (1999). A comparison of measures for detecting natural shapes in cluttered backgrounds. *Int. J. Comput. Vis.* 34, 81–96.

Worley, S. (1996). "A cellular texture basis function," in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY: Association for Computing Machinery), 291–294.

frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 08 January 2024
DOI 10.3389/fcomp.2023.1142621

OPEN ACCESS

EDITED BY
James Elder,
York University, Canada

REVIEWED BY
Carlos Vazquez,
École de technologie supérieure (ÉTS), Canada
Johannes Burge,
University of Pennsylvania, United States

*CORRESPONDENCE
Maria Virginia Bolelli
✉ maria.bolelli2@unibo.it

RECEIVED 11 January 2023
ACCEPTED 06 December 2023
PUBLISHED 08 January 2024

CITATION
Bolelli MV, Citti G, Sarti A and Zucker SW (2024) Good continuation in 3D: the neurogeometry of stereo vision. *Front. Comput. Sci.* 5:1142621. doi: 10.3389/fcomp.2023.1142621

Good continuation in 3D: the neurogeometry of stereo vision

Maria Virginia Bolelli[1,2]*, Giovanna Citti[1], Alessandro Sarti[2] and Steven W. Zucker[3]

[1]Department of Mathematics, University of Bologna, Bologna, Italy, [2]Centre d'analyse et de mathématique sociales (CAMS), CNRS, École des hautes études en sciences sociales (EHESS), Paris, France, [3]Departments of Computer Science and Biomedical Engineering, Yale University, New Haven, CT, United States

Classical good continuation for image curves is based on 2*D* position and orientation. It is supported by the columnar organization of cortex, by psychophysical experiments, and by rich models of (differential) geometry. Here, we extend good continuation to stereo by introducing a neurogeometric model to abstract cortical organization. Our model clarifies which aspects of the projected scene geometry are relevant to neural connections. The model utilizes parameterizations that integrate spatial and orientation disparities, and provides insight into the psychophysics of stereo by yielding a well-defined 3*D* association field. In sum, the model illustrates how good continuation in the (3D) world generalizes good continuation in the (2D) plane.

KEYWORDS

stereo vision, sub-Riemannian geometry, 3D space of position-orientation, 3D association field, neurogeometry

1 Introduction

Binocular vision is the ability of the visual system to provide information about the three-dimensional environment starting from two-dimensional retinal images. Disparities are among the main cues for depth perception and stereo vision but, in order to extract them, the brain needs to determine which features coming from the right eye correspond to those from the left eye, and which do not. This generates a coupling problem, which is usually referred to as the *stereo correspondence problem*. Viewed in the large, stereo correspondence must be consistent with stereo perception more generally, and knowing the relevant features is key for both issues. In this paper we develop an approach to stereo based on the functional organization of the visual cortex, and we identify the geometric features extracted by the binocular cells. This model will be able to extend the notion of good continuation for planar curves to that for 3*D* spatial curves. A simple example demonstrates their application in computing stereo correspondence.

Good continuation in the plane (retinotopic coordinates) is one of the foundational principles of Gestalt perceptual organization. It enjoys an extensive history (Wagemans et al., 2012). It is supported by psychophysical investigations (e.g., Field et al., 1993; Geisler et al., 2001; Elder and Goldberg, 2002; Hess et al., 2003; Lawlor and Zucker, 2013), which reveal connections to contour statistics; it is supported by physiology (orientation selectivity), which reveals the role for long-range horizontal connections (Bosking et al., 1997); and it is supported by computational modeling (Ben-Shahar and Zucker, 2004; Sarti et al., 2007), which reveals a key role for geometry. The notion of orientation underlies all three of these aspects: neurons in visual cortex are selective for orientations, pairs of dots in grouping experiments indicate an orientation, and edge elements in natural images are oriented and related to image statistics. Orientation in space involves two angles, which we shall exploit. Nevertheless, good

continuation in depth is much less well-developed than good continuation in the plane, despite having comparable historical origins. (Koffka, 1963, p. 161-162):

> ...a perspective drawing, even when viewed monocularly, does not give the same vivid impression of depth as the same drawing if viewed through a stereoscope with binocular parallax... for in the stereoscope the tri-dimensional force of the parallax co-operates with the other tri-dimensional forces of organization; instead of conflict between forces, stereoscopic vision introduces mutual reinforcement.

Our specific goal in this paper is to develop a neurogeometrical model of stereo vision, based on the functionality of binocular cells. The main application will be a good continuation model in three dimensions that is analogous to the models of contour organization in two dimensions. We will develop *ad hoc* mathematical instruments, supported by a number of neural and psychophysical investigations (Malach et al., 1993; Uttal, 2013; Deas and Wilcox, 2014, 2015; Khuu et al., 2016; Scholl et al., 2022).

Although only one dimension higher than contours in the plane, contours extending in depth raise subtle new issues; this is why a geometric model can be instructive. First among the issues is the choice of coordinates which, of course, requires a mathematical framework for specifying them. In the plane, position and orientation are natural; smoothness is captured by curvature or the relationship between nearby orientations along a contour. For stereo, there is monocular structure in the left eye and in the right. Spatial disparity is a standard variable relating them, and it is well-known that primate visual systems represent this variable directly (Poggio, 1995). Spatial disparity is clearly a potential coordinate. However, other physiological aspects are less clear. The columnar architecture so powerful for contour organization in the plane is not only monocular: the presence of columns for spatial disparity of binocular cells has been experimentally described in V2 (Ts'o et al., 2009). However, orientation disparity does not seem to be coded in the cortex (see next section). Nevertheless, orientation-selective cells provide the input for stereo so, at a minimum, both position disparity and orientation – one orientation for the right eye and (possibly) another for the left – should be involved. While it is traditional to assume only "like" orientations are matched (Hubel and Wiesel, 1962; Nelson et al., 1977; Marr and Poggio, 1979; Bridge and Cumming, 2001; Chang et al., 2020), our sensitivity to orientation disparity questions this, making orientation disparity another putative variable. We shall show that orientations do play a deep role in stereo, but that it is not necessarily efficient to represent them as a disparity. Furthermore, there is a debate in stereo psychophysics about orientation: since its physiological realization could be confounded with disparity gradients (Mitchison and McKee, 1990; Cagenello and Rogers, 1993), orientation may be redundant. This is not the case, since it is the orientation of the "gradient" that matters. Thus we provide a representation of the geometry of spatial disparity and orientation in support of using good continuation in a manner that both incorporates the biological "givens" and provides a rigorous foundation for the correspondence problem. As has been the case with curve organization, we further believe that our modeling will illuminate the underlying functional architecture for stereo.

Hubel and Wiesel reported disparity-tuned neurons in early, classic work (Hubel and Wiesel, 1970). They observed that single units could be driven from both eyes and that it was possible to plot separate receptive fields (RF) for each eye. We emphasize that these monocular receptive fields are tuned to orientation (Cumming and DeAngelis, 2001; Parker et al., 2016), and a review of neural models can be found in Read (2015).

The classical model for expressing the left/right-eye receptive field combination is the *binocular energy model (BEM)*, first introduced in Anzai et al. (1999b). It encodes disparities through the receptive profiles of simple cells, raising the possibility of both position and phase disparities (Jaeger and Ranu, 2015). However, Read and Cumming (2007), building upon (Anzai et al., 1999a), showed that phase disparity neurons tend to be strongly activated by false correspondence pairs. Other approaches are based on the statistics of natural images (Burge and Geisler, 2014; Jaini and Burge, 2017; Burge, 2020) utilized in an optimal fashion; these lead to more refined receptive field models. Nevertheless, the orientation differences between the two eyes (Nelson et al., 1977), or orientation disparity, should not be neglected. Although there were attempts to incorporate it (Bridge et al., 2001) in energy models, they are limited. The geometrical model we will present incorporates orientation differences directly.

Many other mathematical models for stereo vision based on neural models have been developed. Some claim (e.g., Marr and Poggio, 1979) that orientations should match between the two eyes, although small differences are allowed. This, of course, assumes the structure is frontal-parallel. Subsequently, Jones and Malik (1991) used a set of linear filters tuned to different orientations (and scales) but their algorithm was not built on a neurophysiological basis. Alibhai and Zucker (2000), Li and Zucker (2003), and Zucker (2014) built a more biologically-inspired model that addressed the connections between neurons. Their differential-geometry model employed position, orientations and curvatures in $2D$ retinal planes, modeling binocular neurons with orientations given by tangent vectors of Frenet geometry. Our results here are related, although the geometry is deeper (We develop this below.). A more recent work, based on differential and Riemannian geometry, is developed in Neilson et al. (2018). Before specifying these results, however, we introduce the specific type of geometry that we shall be using. It follows directly from the columnar organization often seen in predators and primates.

1.1 Columnar architectures and sub-Riemannian geometry

We propose a sub-Riemannian model for the cortical-inspired geometry underlying stereo vision based on the encoding of positional disparities and orientation differences in the information coming from the two eyes. We build on neuromathematical models, starting from the work of Koenderink and van Doorn (1987) and Hoffman (1989), with particular emphasis on the neurogeometry of monocular simple cells (Petitot and Tondut, 1999; Citti and Sarti, 2006; Sarti et al., 2007; Petitot, 2008; Sanguinetti et al., 2010; Sarti and Citti, 2015; Baspinar et al., 2020).

To motivate our mathematical approach, it is instructive to build on an abstraction of visual cortex. We start with monocular information, segregated into ocular dominance bands (LeVay et al., 1975) in layer 4; these neurons have processes that extend into the superficial layers. We cartoon this in Figure 1, which shows an array of orientation hypercolumns arranged over retinotopic position. It is colored by dominant eye inputs; the binocularly-driven cells tend to be closer to the ocular dominance boundaries, while the monocular cells are toward the centers. A zoom emphasizes the orientation distribution along a few of the columns near each position; horizontal connections (not shown) effect the interactions between these units. This raises the basic question in this paper: *what is the nature of the interaction among groups of cells representing different orientations at nearby positions and innervated by inputs from the left and right eyes?* The physiology suggests (Figure 1C) the answer lies in the interactions among both monocular and binocular cells; our model specifies this interaction, starting from the monocular ones and building analogously into a columnar organization.

1.2 Informal setup and overview

Since much of the paper is technical, we here specify, informally, the main ingredients of the model and the results. We first list several of the key points, then illustrate them directly.

- Stereo geometry enjoys a mathematical structure that is a formal extension of plane curve geometry. In the plane, points belonging to a curve are described by an orientation at a position, and these are naturally represented as elements (orientation, position) of columns. In our model, these become abstract fibers. The collection of fibers across position is a fiber bundle. Elements of the (monocular) fiber can be thought of as neurons.
- Our geometrical model is based on tangents and curvatures. Tangents naturally relate to orientation selectivity, and are commonly identified with "edge" elements in the world. We shall occasionally invoke this relationship, for intuition and convenience, but some caution is required. While edge elements comprising, e.g., a smooth bounding contour are tangents, the converse is not necessarily true (e.g., elongated attached highlights or hair textures). Instead, our model should be viewed as specifying the constraints relevant to understanding neural circuitry; see Section 1.3.
- To elaborate the previous point: the tangents in our model need not be edges in the world; they are neural responses. The constraints in our model can be used to determine whether these responses should be considered as "edges." This is why the model is built from the geometry of idealized space curves: to support such inferences.
- For stereo, we shall need fibers that are a "product" of the left and right-eye monocular columns. This is the reason why we choose position, positional disparity and orientations from the left and right eyes respectively, as the natural variables that describe the stereo fiber over each position. We stress that these fibers are not necessarily explicit in the cortical architecture.
- Curvature provides a kind of "glue" to enable transitions from points on fibers to nearby points on nearby fibers. These transitions specify "integral curves" through the stereo fiber bundle.
- The integral curve viewpoint provides a direction of information flow (information diffuses through the bundle) thereby suggesting underlying circuits.
- The integral curves formalize *association field* models. Their parameters describe the spray of curves that is well in accordance with $3D$ curves as studied in psychophysical experiments in Hess and Field (1995), Hess et al. (1997), and Khuu et al. (2016).
- Our formal theory addresses several conjectures in the literature. The first is the *identity hypothesis* (Kellman et al., 2005a,b) and the organization of units for curve interpolation (cf. Anderson et al., 2002); we show how tangents are natural "units" and how they can be organized. The second concerns the nature of the organization (Li and Zucker, 2006), where we resolve a conjecture regarding the interpolating object (see Proposition 3.2 below).
- Our formal theory provides a new framework for specifying the correspondence problem, by illustrating how good continuation in the 3-D world generalizes good continuation in the 2-D plane. This is the point where consistent binocular-binocular interactions are most important.
- Our formal theory has direct implications for understanding torsional eye movements. It suggests, in particular, that the rotational component is not simply a consequence of development, but that it helps to undo inappropriate orientation disparity changes induced by eye movements. This role for Listing's Law will be treated in a companion paper (in preparation); see also the excellent paper (Schreiber et al., 2008).

We now illustrate these ideas (Figure 2). Consider a three-dimensional stimulus as a space curve $\gamma : \mathbb{R} \longrightarrow \mathbb{R}^3$, with a unit-length tangent at the point of fixation. Since the tangent is the derivative of a curve, the binocular cells naturally encode the unitary tangent direction $\dot{\gamma}$ to the spatial $3D$ stimulus γ. This space tangent projects to a tangent orientation in the left eye[1], and perhaps the same or a different orientation in the right eye. A nearby space tangent projects to another pair of monocular tangents, illustrated as activity in neighboring columns. Note how connections between the binocular neurons support consistency along the space curve. It is this consistency relationship that we capture with our model of the stereo association field.

Since space curves live in 3D, two angles are required to specify its space tangent at a point. In other words, monocular tangent angles span a circle in the plane; space tangent angles span a 2-sphere in 3D. In terms of the projections into the left-eye and the right-eye, the space tangent can be described by the

1 We are here being loose with language. By a tangent orientation in the left eye, we mean the orientation of a left-eye innervated column in V1.

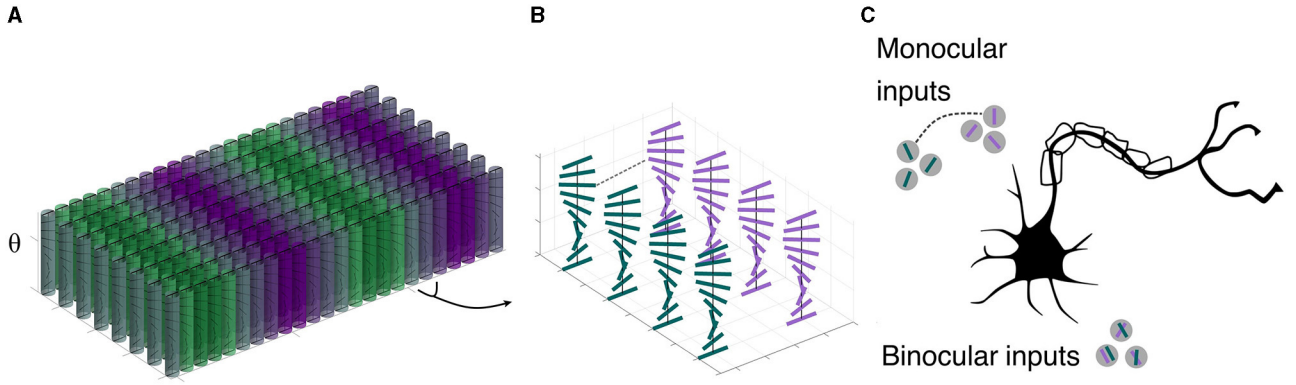


FIGURE 1
Cartoon of visual cortex, V1, superficial layers. **(A)** Macroscopic organization: A number of (abstracted) orientation hypercolumns, colored by left-eye (green)/right-eye (purple) dominant inputs. The color grading emphasizes that at the center of the ocular dominance bands the cells are strongly monocular, while at the boundaries they become binocularly-driven. **(B)** A zoom in to a few orientation columns showing left and right monocular cells at the border of ocular dominance bands. Cells in these nearby columns will provide the anatomical substrate for our model. **(C)** More recent work shows that both monocular and binocular inputs matter to these cells (redrawn from Scholl et al., 2022, using data from ferret). This more advanced wiring suggests the connection structures in our model.

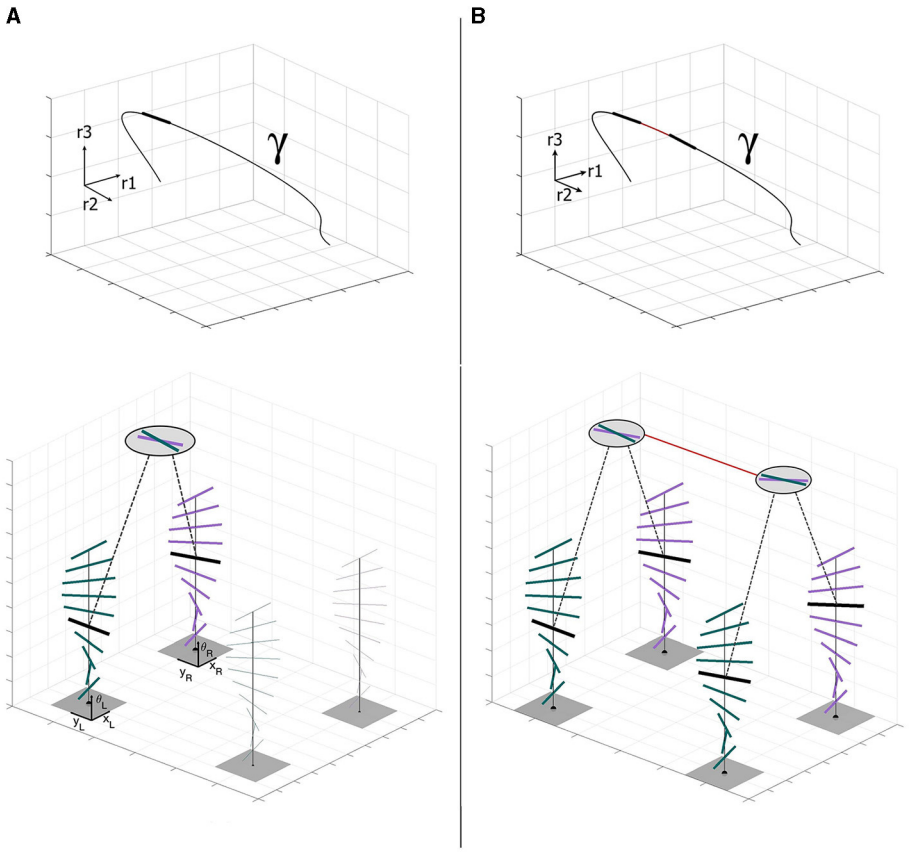


FIGURE 2
(A) Stereo projection of the highlighted tangent vector to the stimulus $\gamma \in \mathbb{R}^3$ in the left-eye innervated and right-eye innervated monocular orientation columns (Each short line denotes a neuron by its orientation preference.). Joint activity across the eyes, which denotes the space tangent, is illustrated by the binocular neuron (circle). Note the two similar but distinct monocular orientations. Connections from the actively stimulated monocular neurons to the binocular neuron are shown as dashed lines. **(B)** Stereo projection of a consecutive pair of tangents to the stimulus $\gamma \in \mathbb{R}^3$ in the left and right retinal columns. Each space tangent projects to a different pair of monocular columns because of the spatial disparity. Consistency in the responses of these four columns corresponds to consistency between the space tangents attached to nearby positions along γ. This consistency is realized through the binocular neural connection (solid line).

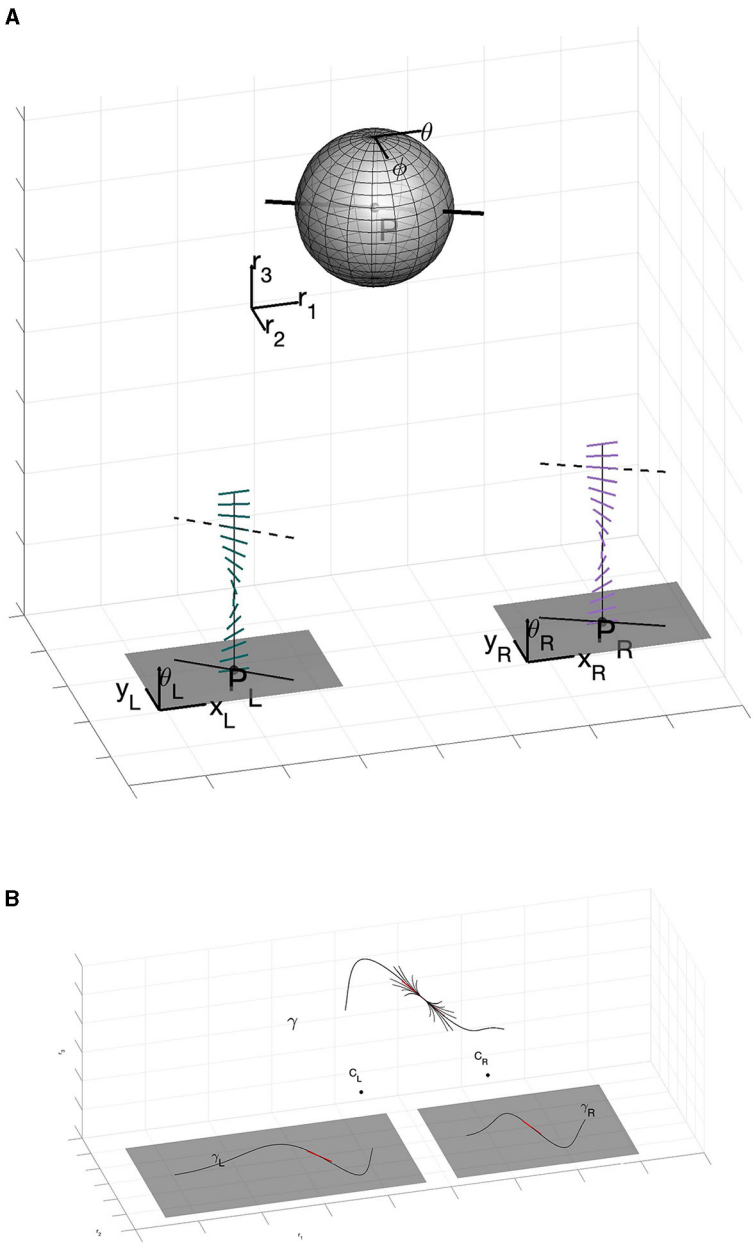


FIGURE 3
(A) The full geometry of stereo. Note how the stereo correspondence problem allows to establish the relationship between the 3*D* tangent point (P, θ, ϕ) and the projections p_L and p_R, the disparity and the orientations θ_L and θ_R. **(B)** Main result of the paper. The three-dimensional space curve γ is enveloped by the 3*D* association field centered at a point. Formally, this association field is a fan of integral curves in the sub-Riemmanian geometry computed entirely within the columnar architecture (It is specifically described by Equation (36) with varying c_1 and c_2 in $\mathbb{R}$, but that will take some work to develop.).

parameters $n = (\theta, \varphi)$ of $\mathbb{S}^2$ (Figure 3A). Thus, we can suitably describe the space of stereo cells – the full set of space tangents at any position in the 3*D* world – as the manifold of positions and orientations $\mathbb{R}^3 \rtimes \mathbb{S}^2$. Moving from one position in space to another, and changing the tangent orientation to the one at the new position, amounts to what is called a *group action* on the appropriate manifold. We informally introduce these notions in the next subsection; a more extensive introduction to these ideas is in Appendix A (Supplementary material).

1.2.1 Sub-Riemannian geometry

We live in a 3*D* world in which distances are familiar; that is, a space of points with a Euclidean distance function defined between any pair of them. Apart from practical considerations we can move in any direction we would like. Cars, however, have much more restricted movement capabilities. They can move forward or backward, but not sideways. To move in a different direction, cars must turn their wheels. Here is the basic analogy: in cortical space information can move to a new retinotopic position in a tangent

direction, or it can move up or down a column (orientation fiber) to change direction. Moving in this fashion, from an orientation at a position to another orientation at a nearby position, is clearly more limited than arbitrary movements in Euclidean space. Euclidean geometry, as above, is an example of a Riemannian geometry; the limitations involved in moving through a cortical columnar space specify a sub-Riemannian geometry (Citti and Sarti, 2014; Citti et al., 2015; Sarti et al., 2019). Just as cars can move along roads that are mostly smooth, excitatory neurons mainly connect to similarly "like" (in orientation) excitatory neurons. This chain of neurons indicates a path through sub-Riemannian space (Agrachev et al., 2019); the fan of such paths is the cortical connectivity which can be considered the neural correlate of association fields. Again, for more information please consult Appendix A (Supplementary material).

Moving now out to the world, we must be able to move between all points. Repeating the above metaphor more technically, we equip $\mathbb{R}^3 \rtimes \mathbb{S}^2$ with a group action of the three-dimensional Euclidean group of rigid motions $SE(3)$. Notice, importantly, that this group is now acting on the product space of positions and orientations. A bit more is required, though, since the geometry of stereo vision is not solved only with these punctual and directional arguments. As we showed in Figure 2 there is the need to take into account the relationships between nearby tangents; in geometric language this involves a suitable type of connections. It is therefore natural to look at integral curves of the sub-Riemannian structure, which encode in their coefficients the fundamental concept of $3D$ curvature and torsion. An example of this is shown in Figure 3B. Notice how the $3D$ association field envelopes a space curve, in the same way that a $2D$ association field envelopes a planar curve. This figure illustrates, in a basic way, the fundamental result in this paper.

1.3 On the neuro-geometric approach

There are many different ways to approach mathematical modeling in vision. One could, for example, ask what is the best an ideal observer could do for the stereo problem working directly on image data (Burge and Geisler, 2014; Burge, 2020). This requires specifying the task, e.g., disparity at a point; a database of images on which the estimation is to be carried out; and a specification of the output. The approach is fundamentally statistical, and has been successful at predicting discrimination thresholds and optimal receptive field designs for patches of natural images. We seek to go the next step – to specify the relationship between receptive fields; i.e., between neurons. Note that the complexity multiplies enormously. At the behavioral level this raises the question of grouping, or determining the combinations of disparities, or Gabor patch samples, that belong together. The complexity arises because this must be evaluated over all possible arrangements of patches, be they along curves, or surfaces, or combinations thereof. In effect, the output specification is pushed toward co-occurrence phenomena, and these toward neural connections.

Our working hypothesis is that there is a deep functional relationship between structure in the brain and structure in the world, and that geometry is the right language with which to capture this relationship, especially as regards connectivity between neurons and their functionality. The *neuro-geometric approach* is precisely this; an attempt to capture how the structure of cortical connectivity (and other functional properties) are reflected in the phenomena of visual perception.

At first blush this might seem completely unrelated to the statistics of natural images, and how these could be informative of neural connections, but we believe that there is a fundamental relationship. Consider, to start, the distribution of oriented edge elements in a small patch. Pairwise edge statistics are well-studied (August and Zucker, 2000; Geisler et al., 2001; Elder and Goldberg, 2002; Sanguinetti et al., 2010), and indicate how orientation changes are distributed over (spatially) nearby edge elements. Co-linear and co-circular patterns emerge from these studies, as well as in third-order statistics of edges (Lawlor and Zucker, 2013). Interestingly, in Singh and Fulvio (2007) and Geisler and Perry (2009) deviation from co-circular behavior emerges.[2] In particular, Geisler and Perry (2009) proposes a parabolic model to explain these statistical evidences, that is consistent with previous results if we consider a composition of the joint action of cocircularity and parallelism cues (as found to factor for example in Elder and Goldberg, 2002). To elaborate, it begins by following either a co-circular or linear term, followed by the composition with another circular or linear term. The outcome of this process is described as a spline-like behavior that can approximate a parabola. In Sanguinetti et al. (2010), it has been shown that the histogram of the co-occurrence of edges in a natural image provides the same probability kernel we could find with geometric analysis instruments. As a result, statistical measurements are integrated into the geometric approach.

The geometric analysis that we shall use is continuous mathematics, and is essentially differential (Tu, 2011). This has important implications. First, the relationships that matter are those over small neighborhoods, not over "long" distances. Thus at a point there is an orientation (tangent) and a curvature. These barely change as one moves a tiny distance from the point. Thus we are not considering (in this paper) what happens behind (relatively) large occluders, when longer distances devoid of intermediate structure separate structure (Singh and Fulvio, 2005; Fulvio et al., 2008). Such problems are important but are outside the scope of this paper. Second, because the mathematics is continuous, we shall not consider sampling issues (Warren et al., 2002). To the extent that it matters, we shall assume discrete entities are sufficiently densely distributed that they function as if they were continuous (Zucker and Davis, 1988). In this sense our analysis is restricted to early vision. It does not necessarily account for the full range of cognitive tasks, which may well invoke higher-order computations over longer distances and even richer abstractions.

It has been observed that edge statistics for curves in the world depart from co-circularity. To quote (Geisler and Perry, 2009): "Except for a direction of zero, where the orientation difference is consistent with a collinear relationship, the highest-likelihood orientation differences are less than those predicted by a co-circular

2 Of course we need to take into account the difficulty of measurements of coupled position-orientation variables for small difference of angle and position. This is due to the well-known intrinsic uncertainty of measurement in the non-commutative group of position and orientation (Barbieri et al., 2012).

relationship." We believe this has to do with the notion of curvature used: whether it is purely local, or an estimate over distance. In summary of the geometric approach, we explain this as follows. Our constraints can be used in two rather different ways. First, from a computational perspective, one can "integrate" the local constraints into more global objects. This is the approach used in the example section of this paper, and could give rise to an "average" curvature over some distance. The second approach is more distributed, and may be closer to a neurobiological implementation (Ben-Shahar and Zucker, 2004). In this second approach (not developed in this paper, but see Li and Zucker (2006) and Figure 6), the local computations overlap to enforce consistency (The scale of such computations would be a small factor larger than that indicated in Figure 2B). This scale corresponds to the extent of biological "long range horizontal connections" but is smaller than many of the occluders used in psychophysical experiments. In other words, to emphasize this distinction, in the former case the use of integral curves may be closer to the parabolic relations observed in scene statistics (Geisler and Perry, 2009).[3] Our use of the term "co-circularity" is in the latter sense.

1.4 Overview of paper

The paper is organized as follows: in Section 2, we describe the geometrical and neuro-mathematical background underlying the problem of stereo vision. In particular, we review the standard stereo triangulation technique to relate the coordinate system of one retina with the other, and put them together in order to reconstruct the three-dimensional space. Then, we briefly review the classical neurogeometry of monocular simple cells selective for orientation and the underlying connections. The generalization of approximate co-circularity for stereo is also introduced. In Section 3, starting from binocular receptive profiles, we introduce the neuro-mathematical model for binocular cells. First we present the cortical fiber bundle of binocular cells. It follows the differential interpretation of the binocular profiles in terms of the neurogeometry of the simple cells, and we show how this is well in accordance with the results of the stereo triangulation. Then, we give a mathematical definition of the manifold $\mathbb{R}^3 \rtimes \mathbb{S}^2$ with the sub-Riemannian structure. Finally, we study the integral curves and the suitable change of variables that allow us to switch our analysis from cortical to external space. In Section 4 we proceed to the validation of our geometry with respect to psychophysical experiments. We combine information about the psychophysics of $3D$ perception and formal conjectures; it is here that we formulate a $3D$ association field analogous to the $2D$ association field. At the end, we show an example of a representation of a stimulus (from image planes to the full 3D and orientation geometry) and how our integral curves properly connect corresponding points. This illustrates the use of our model as a basis for solving the correspondence problem.[4]

3 The crucial point is that the curves demonstrate locally quadratic (not linear) behavior.

4 Portions of this material were presented at Bolelli et al. (2023a).

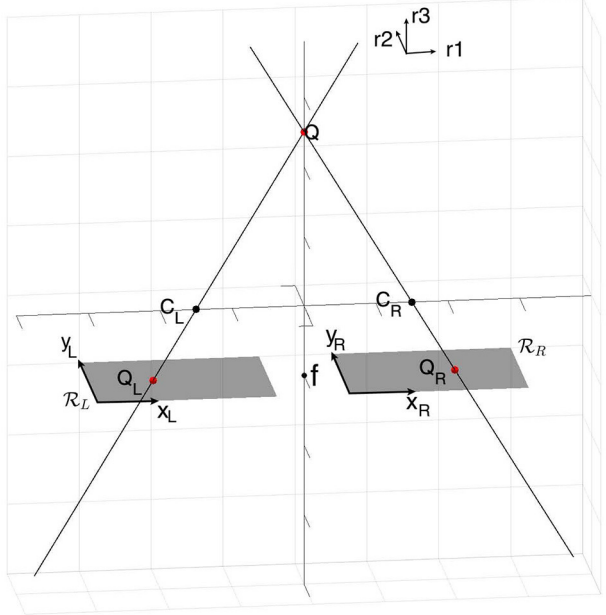


FIGURE 4
Reconstruction of the $3D$ space point Q through points Q_L the retinal plane $\mathcal{R}_L$ and Q_R in $\mathcal{R}_R$.

2 Stereo vision and neuro-mathematical background

2.1 Stereo geometry

In this subsection, we briefly recall the geometrical configuration underlying $3D$ vision, to define the variables that we use in the rest of the paper, mainly referring to (Faugeras, 1993, Ch. 6). For a complete historical background see Howard (2012); Howard and Rogers (1995).

2.1.1 Stereo variables

We consider the global reference system (O, i, j, k) in $\mathbb{R}^3$, with $O = (0, 0, 0)$, and coordinates (r_1, r_2, r_3). We introduce the optical centers $C_L = (-c, 0, 0)$ and $C_R = (c, 0, 0)$, with c real positive element, and we define two reference systems: (C_L, i_L, j_L), (C_R, i_R, j_R), the reference systems of the retinal planes $\mathcal{R}_L$ and $\mathcal{R}_R$ with coordinates respectively (x_L, y), (x_R, y). In the global system we suppose the retinal planes to be parallel and to have equation $r_3 = f$, with f denoting the focal length. This geometrical set-up is shown in Figure 4.

Remark 2.1. If we know the coordinate of a point $Q = (r_1, r_2, r_3)^T$ in $\mathbb{R}^3$, then it is easy to project it in the two planes via perspective projection, having c the coordinate of the optical centers and f focal length. This computation defines two projective maps Π_L and Π_R, respectively, for the left and right retinal

planes:

$$\Pi_L : \mathbb{R}^3 \longrightarrow \mathbb{R}^2 \qquad \Pi_R : \mathbb{R}^3 \longrightarrow \mathbb{R}^2$$
$$\begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix} \mapsto \begin{pmatrix} \frac{f(r_1+c)}{r_3} \\ \frac{fr_2}{r_3} \end{pmatrix}, \qquad \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix} \mapsto \begin{pmatrix} \frac{f(r_1-c)}{r_3} \\ \frac{fr_2}{r_3} \end{pmatrix}. \tag{1}$$

Proof. A point on the left retinal plane of local coordinates $(x_L, y)^T$ has global coordinates $Q_L = (-c + x_L, y, f)^T$, and it corresponds to a point $Q = (r_1, r_2, r_3)^T$ in the Euclidean $\mathbb{R}^3$ such that C_L, Q_L and Q are aligned. This means that the vectors $Q_L - C_L = (x_L, y, f)^T$ and $Q - C_L = (r_1 + c, r_2, r_3)^T$ are parallel, obtaining the following relationships:

$$x_L = f\frac{r_1 + c}{r_3}, \quad y = f\frac{r_2}{r_3}. \tag{2}$$

Analogously, considering Q_R and C_R, we get:

$$x_R = f\frac{r_1 - c}{r_3}, \quad y = f\frac{r_2}{r_3}. \tag{3}$$

□

In a standard way, the *horizontal disparity* is defined as the differences between retinal coordinates

$$d := \frac{x_L - x_R}{2}, \tag{4}$$

up to a scalar factor. Moreover, it is also possible to define the coordinate x as the average of the two retinal coordinates $x := \frac{x_L + x_R}{2}$, leading to the following change of variables:

$$\begin{cases} x = \frac{fr_1}{r_3} \\ y = \frac{fr_2}{r_3} \\ d = \frac{fc}{r_3} \end{cases} \longleftrightarrow \begin{cases} r_1 = \frac{xc}{d} \\ r_2 = \frac{yc}{d} \\ r_3 = \frac{fc}{d} \end{cases}, \tag{5}$$

where the set of coordinates (x, y, d) is known as *cyclopean coordinates* (Julesz, 1971).

2.1.2 Tangent estimation

Corresponding points in the retinal planes allow to project back into $\mathbb{R}^3$. An analogous reasoning can be done for the tangent structure: if we have tangent vectors of corresponding curves in the retinal planes, it is possible to project back and recover an estimate of the $3D$ tangent vector. Let us recall here this result; a detailed explanation can be found in Faugeras (1993).

Remark 2.2. Let γ_L and γ_R be corresponding left and right retinal curves; i.e., perspective projections of a curve $\gamma \in \mathbb{R}^3$ through optical centers C_L and C_R with focal length f. Knowing the left and right retinal tangent structures, it is possible to recover the direction of the tangent vector $\dot{\gamma}$.

Proof. Starting from a curve $\gamma \in \mathbb{R}^3$, we project it in the two retinal planes obtaining $\gamma_L = \Pi_L(\gamma)$ and $\gamma_R = \Pi_R(\gamma)$ from Equation (1). The retinal tangent vectors are obtained through the Jacobian matrix[5] of the left and right retinal projections $\dot{\gamma}_{L,R}(t) = (J_{\Pi_{L,R}})_{\gamma(t)}\dot{\gamma}(t)$:

$$\dot{\gamma}_R(t) = \begin{pmatrix} \frac{f(\gamma_3\dot{\gamma}_1 + (c-\gamma_1)\dot{\gamma}_3)}{\gamma_3(t)^2} \\ \frac{f(\gamma_3\dot{\gamma}_2 - \gamma_2\dot{\gamma}_3)}{\gamma_3^2} \end{pmatrix}, \dot{\gamma}_L(t) = \begin{pmatrix} \frac{f(\gamma_3\dot{\gamma}_1 - (c+\gamma_1)\dot{\gamma}_3)}{\gamma_3(t)^2} \\ \frac{f(\gamma_3\dot{\gamma}_2 - \gamma_2\dot{\gamma}_3)}{\gamma_3^2} \end{pmatrix}. \tag{6}$$

Extending the tangent vectors and the points into $\mathbb{R}^3$, we get $\tilde{t}_L = (\dot{\gamma}_{L1}, \dot{\gamma}_{L2}, 0)^T$, and $\tilde{m}_L = (\gamma_{L1} - c, \gamma_{L2}, f)^T$, and $U_{t_L} = (P_L)^{-1}\tilde{m}_L \times (P_L^{-1})\tilde{t}_L$, with the projection matrix $P_L = \begin{pmatrix} 1 & 0 & -c/f \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. The same reasoning holds for the right structure, with projection matrix $P_R = \begin{pmatrix} 1 & 0 & c/f \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

Then $U_{t_R} \times U_{t_L}$ is a vector parallel to the tangent vector $\dot{\gamma}$:

$$\begin{aligned} U_{t_R} \times U_{t_L} &= \begin{pmatrix} \underbrace{\frac{f^4 2c(\dot{\gamma}_2\gamma_3 - \dot{\gamma}_3\gamma_2)}{\gamma_3^4}}_{\lambda(t)}\dot{\gamma}_1, \\ \frac{f^4 2c(\dot{\gamma}_2\gamma_3 - \dot{\gamma}_3\gamma_2)}{\gamma_3^4}\dot{\gamma}_2, \frac{f^4 2c(\dot{\gamma}_2\gamma_3 - \dot{\gamma}_3\gamma_2)}{\gamma_3^4}\dot{\gamma}_3 \end{pmatrix}^T \\ &= \lambda(t)\left(\dot{\gamma}_1(t), \dot{\gamma}_2(t), \dot{\gamma}_3(t)\right)^T \\ &= \lambda(t)\dot{\gamma}(t). \end{aligned} \tag{7}$$

□

Although this section has been based on the geometry of space curves and their projections, we observe that related geometric approaches have been developed for planar patches and surfaces; see, e.g., Li and Zucker, 2008; Oluk et al., 2022 and references therein.

2.2 Elements of neurogeometry

We now provide background on the geometric modeling of the monocular system, and good continuation in the plane. Our goal is to illustrate the role of sub-Riemannian geometry in the monocular system, which will serve as the basis for generalization to the stereo system.

2.2.1 Classical neurogeometry of simple cells

We model the activation map of a cortical neuron's receptive field (RF) by its receptive profile (RP) φ. A classical example is the receptive profiles of simple cells in V1, centered at position

5 The Jacobian matrix $(J_\Pi)_p$ evaluated at point p represents how to project displacement vectors (in the sense of derivatives or velocities or directions). In details, if $\dot{\gamma}(t)$ is the displacement vector in $\mathbb{R}^3$, then the matrix product $(J_\Pi)_{\gamma(t)}\dot{\gamma}(t)$ is another displacement vector, but in $\mathbb{R}^2$. In other words, the Jacobian matrix is the differential of Π at every point where Π is differentiable; common notation includes J_Π or $D\Pi$.

(x, y) and orientation θ, modeled (e.g., in Daugman, 1985; Jones and Palmer, 1987; Barbieri et al., 2014b) as a bank of Gabor filters $\varphi_{\{x,y,\theta\}}$. RPs are mathematical models of receptive fields; they are operators which act on a visual stimulus.

Formally, it is possible to abstract the primary visual cortex as $\mathbb{R}^2 \times \mathbb{S}^1$, or position-orientation space, thereby naturally encoding the Hubel/Wiesel hypercolumnar structure (Hubel and Wiesel, 1962). An example of this structure is displayed in Figure 5D from Ben-Shahar and Zucker (2004).

Following the model of Citti and Sarti (2006), the set of simple cells' RPs can be obtained via translations along a vector $(x, y)^T$ and rotation around angle θ from a unique "mother" profile $\varphi_0(\xi, \eta)$:

$$\varphi_0(\xi, \eta) = \exp\left(\frac{2\pi i\xi}{\lambda}\right) \exp\left(-\frac{\xi^2 + \eta^2}{2\sigma^2}\right), \tag{8}$$

This RP is a Gabor function with even real part and odd imaginary part (Figure 7A). Translations and rotations can be expressed as:

$$T_{(x,y,\theta)}(\xi, \eta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix} + \begin{pmatrix} x \\ y \end{pmatrix}, \tag{9}$$

where $T_{(x,y,\theta)}$ denotes the action of the group of rotations and translations $SE(2)$ on $\mathbb{R}^2$. This group operation associates to every point (ξ, η) a new point $(\tilde{x}, \tilde{y})$, according to the law $(\tilde{x}, \tilde{y}) = T_{(x,y,\theta)}(\xi, \eta)$. Hence, a general RP can be expressed as

$$\varphi_{(x,y,\theta)}(\xi, \eta) = \varphi_0(T^{-1}_{(x,y,\theta)}(\xi, \eta)), \tag{10}$$

and this represents the action of the group $SE(2)$ on the set of receptive profiles.

The retinal plane $\mathcal{R}$ is identified with the $\mathbb{R}^2$ plane, whose coordinates are (x, y). When a visual stimulus $I : \mathcal{R} \longrightarrow \mathbb{R}^+$ of intensity $I(x, y)$ activates the retinal layer, the neurons centered at every point (x, y) produce an output $O(x, y, \theta)$, modeled as the integral of the signal I with the set of Gabor filters:

$$O(x, y, \theta) = \int_{\mathcal{R}} \varphi_{\{x,y,\theta\}}(\xi, \eta) I(\xi, \eta) d\xi d\eta, \tag{11}$$

where the function I represents the retinal image.

For (x, y) fixed, we will denote $\bar{\theta}$ the point of maximal response:

$$\max_{\theta} |O(x, y, \theta)| = |O(x, y, \bar{\theta})|. \tag{12}$$

We will then say that the point (x, y) is lifted to the point $(x, y, \bar{\theta})$. This is extremely important conceptually to understand our geometry: it illustrates how an image point, evaluated against a simple cell RP, is lifted to a "cortical" point by introducing the orientation explicitly. If all the points of the image are lifted in the same way, the level lines of the $2D$ image I are lifted to new curves in the $3D$ cortical space $(x, y, \bar{\theta})$.

We shall now recall a model of the long range connectivity which allows propagation of the visual signal from one cell in a column to another cell in a nearby column. This is formalized as a set of directions for moving in the cortical space $(x, y, \bar{\theta})$, in the sense of vector fields. This is important because it will be necessary to move within this space, across both positions and orientations.

To begin, in the right hand side of the Equation (11) the integral of the signal with the real and imaginary part of the Gabor filter is expressed. The two families of cells have different shapes, hence they detect (or play a role in detecting) different features. Since the odd-symmetry cells suggest boundary detection, we concentrate on them, but this is a mathematical simplification. The output of a simple cell can then be locally approximated as $O(x, y, \theta) = -X_{3,p}(I_\sigma)(x, y)$, where $p = (x, y, \theta) \in SE(2)$, I_σ is a smoothed version of I, obtained by convolving it with a Gaussian kernel, and

$$X_{3,p} = -\sin\theta \partial_x + \cos\theta \partial_y, \tag{13}$$

is the directional derivative in the direction $\vec{X}_{3,p} = (-\sin\theta, \cos\theta, 0)^T$. From now on, we will denote (by a slight abuse of notation) $\omega^\star := \vec{X}_{3,p}$ to remind the reader familiar with the language of 1-forms the correspondence of these quantities, and the relation with the Hodge star operator.[6]

Now, think of vector fields as defining a coordinate system at each point in cortical space. Then, in addition to above, the vector fields orthogonal to $X_{3,p}$ are:

$$X_{1,p} = \cos\theta \partial_x + \sin\theta \partial_y, \quad X_{2,p} = \partial_\theta \tag{14}$$

and they define a 2-dimensional admissible tangent bundle[7] to $\mathbb{R}^2 \times \mathbb{S}^1$. One can define a scalar product on this space by imposing the orthonormality of $X_{1,p}$ and $X_{2,p}$: this determines a sub-Riemannian structure on $\mathbb{R}^2 \times \mathbb{S}^1$.

The visual signal propagates, in an anisotropic way, along cortical connectivity and connects more strongly cells with comparable orientations. This propagation has been expressed by the geometry just developed and 2-dimensional contour integration. This is the neural explanation of the Gestalt law of good continuation (Koffka, 1963; Kohler, 1967). It can be directly expressed as co-circularity in the plane (Parent and Zucker, 1989), to describe the consistency and the compatibility of neighboring oriented points, in accordance with specific values of curvature. An example of these compatibilities can be found in Figure 5A. It is complemented by psychophysical experiments, e.g., Uttal, 1983; Smits and Vos, 1987; Ivry et al., 1989. In particular, Field et al. (1993) describe the association rules for 2-dimensional contour integration, introducing the concept of *association fields*. A representation of these connections can be found in Figures 5B, C. Note that this is equivalent to the union (over curvature) in Parent and Zucker (1989). Neurophysiological studies (Blasdel, 1992; Malach et al., 1993; Bosking et al., 1997; Schmidt et al., 1997; Hess et al., 2014) suggest that the cortical correlate of the association field is the long-range horizontal connectivity among cells of similar (but not necessarily identical) orientation preference.

Based on these findings, Citti and Sarti (2006) modeled cortical propagation as propagation along integral curves of the vector fields X_1 and X_2, namely curves $\gamma : [0, T] \subset \mathbb{R} \longrightarrow \mathbb{R}^2 \times \mathbb{S}^1$ described by the following differential equation:

$$\dot{\gamma}(t) = \vec{X}_{1,\gamma(t)} + k\vec{X}_{2,\gamma(t)}, \quad t \in [0, T], \tag{15}$$

6 The purpose of introducing this notation is also to motivate an implication of the mathematical model in Citti and Sarti (2006); see Appendix B.2.1 (Supplementary material) for explanation.

7 as defined in Appendix A3 (Supplementary material).

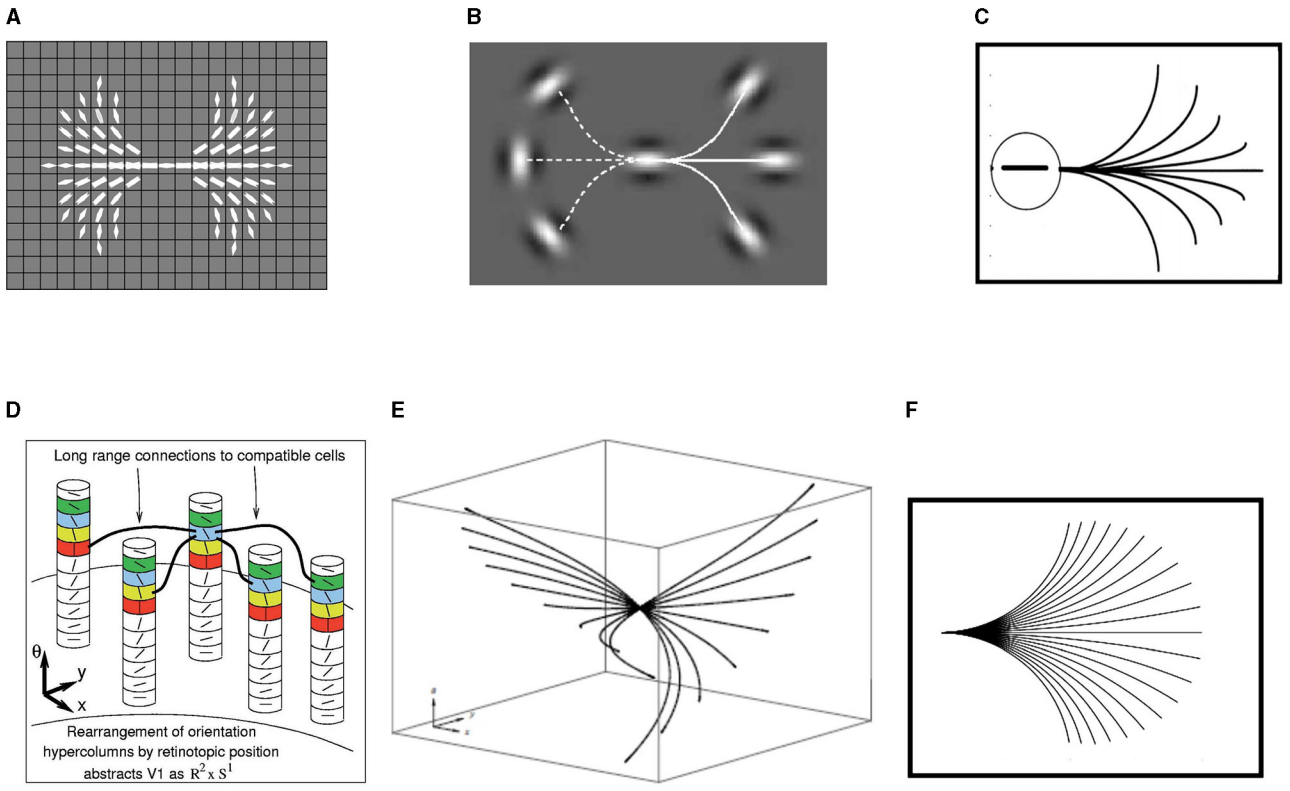


FIGURE 5
(A) Examples of the compatibilities around the central point of the image, derived from planar co-circularity. Brightness encodes compatibility values. Figure adapted from Ben-Shahar and Zucker (2004). **(B)** Starting from the central initial oriented point, the solid line indicates a configuration between the patches where the association exists while the dashed line indicates a configuration where it does not. Figure adapted from Field et al. (1993). **(C)** Association field of Field, Hayes, and Hess. Figure adapted from Field et al. (1993). **(D)** Orientation columns of cells in (x, y, θ) coordinates. Long-range horizontal connections between cells relate an orientation signal at position (x, y, θ) to another orientation at (x', y', θ'). Figure adapted from Ben-Shahar and Zucker (2004). **(E)** Horizontal integral curves in $\mathbb{R}^2 \times \mathbb{S}^1$ generated by the sub-Riemannian model (Citti and Sarti, 2006). **(F)** Projection of the fan of the integral curves in the (x, y) plane. Figure adapted from Citti and Sarti (2006).

obtained by varying the parameter $k \in \mathbb{R}$. (k acts analogously as curvature.) An example of these curves is in Figure 5E. Their $2D$ projection is a close approximation of the association fields (Figure 5F).

A related model has been proposed by Duits et al. (2013). They study the geodesics of the sub-Riemannian structure to take into account all appropriate end-conditions of association fields.

2.2.2 Generalizing co-circularity for stereo

The concept of co-circularity in $\mathbb{R}^2$ has been developed by observing that a bidimensional curve γ can be locally approximated at 0 via the osculating circle.[8] Alibhai and Zucker (2000), Li and Zucker (2003), and Li and Zucker (2006) generalize this concept with the Frenet differential geometry of a three dimensional curve.

While in the two-dimensional case the approximation of the curve using the Frenet $2D$ basis causes the curvature to appear in the coefficient of the Taylor series development ($1st$ order), in the three-dimensional case the coefficients involve both the curvature and torsion. So, in Alibhai and Zucker (2000) the authors propose heuristically to generalize the osculating circle for space curves with an osculating helix, with a preference for r_3-helices to improve stability in terms of camera calibration. In this way the orientation disparity is encoded in the behavior of the helix in the $3D$ space: there is no difference in orientation in the retinal planes if the helix is confined to be in the fronto-parallel plane (the helix becomes a circle); otherwise moving along the $3D$ curves the retinal projections have different orientations.

In Li and Zucker (2003, 2006) they observe that, by introducing the curvature variable as a feature in the two monocular structures, and assuming correspondence, it is possible to reconstruct the $3D$ Frenet geometry of the curve, starting from the two-dimensional Frenet geometry, up to the torsion parameter. In particular, they prove:

Proposition 2.1. Given two perspective views of a $3D$ space curve with full calibration, the normal N and curvature k at a curve

8 Locally, a curve can be approximated by its osculating circle and, at a slightly larger scale, by the integral (parabolic) curve through the first two Taylor terms. The first approximation is co-circularity; the second is a parabolic curve. The second is an accurate model over large distances; see discussion in Section 1.3. However, since in this paper we are working over small distances and with cortical sampling (Figure 2), there is essentially no difference between them; see Figure 22.4 in Zucker (2006) and Sanguinetti et al. (2010) for a direct comparison.

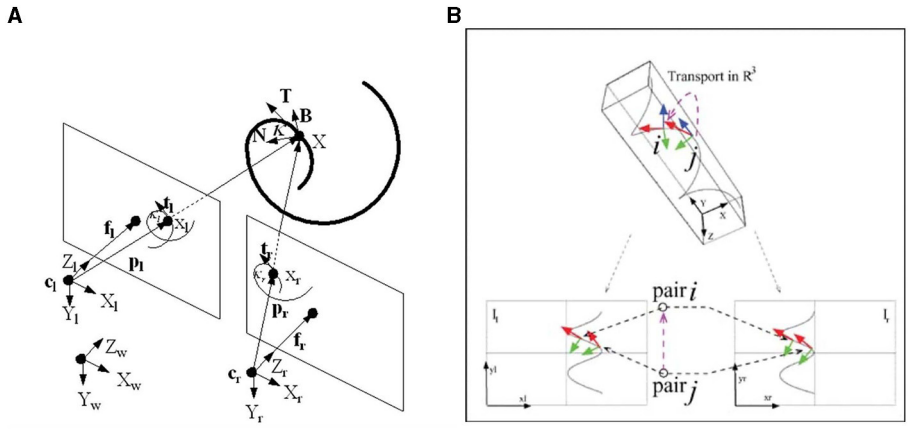


FIGURE 6
(A) Geometrical setup of Proposition 2.1. The spiral curve 3*D* projects in the left and right retinal planes together with the Frenet structure. **(B)** Stereo correspondence between pairs of (left-right) pairs of tangents. Both figures are taken from Li and Zucker (2006).

space point are uniquely determined from the positions, tangents, and curvatures of its projections in two images. Thus the Frenet frame $\{T, N, B\}$ and curvature k at the space point can be uniquely determined.

Hence, using the knowledge of the Frenet basis together with the fundamental addition of the curvature variable, Zucker et al. applied the concept of *transport*. This allowed moving the 3*D* Frenet frame in a consistent way with the corresponding 2*D* Frenet structures of the left and right retinal planes, to establish stereo correspondence between pairs of (left and right) pairs of tangents (see Figure 6B).

Remark 2.3. The model that we propose in this paper is related to, but differs from, what has just been stated. In particular, to remain directly compatible with the previous neuro-geometric model, we will work only with the monocular variables of position and orientation. Rather than using curvature directly, we shall assume that these variables are encoded within the connections; mathematically they appear as parameters. A theoretical result of our model is that the heuristic assumption regarding the r_3-helix can now be established rigorously.

Let us also mention the paper (Abbasi-Sureshjani et al., 2017), where the curvature was considered as independent variable and helices have been obtained in the 2*D* space.

3 The neuromathematical model for stereo vision

Here, we do not want to directly impose a co-circularity property: our scope is to model the behavior of binocular cells, and deduce properties of propagation, which will ultimately induce a geometry of 3D good continuation laws.

3.1 Binocular profiles

Binocular neurons receive inputs from both the left and right eyes. To facilitate calculations, we assume these inputs are first combined in simple cells in the primary visual cortex, a widely studied approach (Anzai et al., 1999b; Cumming and DeAngelis, 2001; Menz and Freeman, 2004; Kato et al., 2016). It provides a first approximation in which binocular RPs are described as the product of monocular RPs; see Figure 7. This model is clearly an oversimplification, in several senses. First, it leaves out the more refined receptive fields discussed in Section 1.3. Second, it leaves out the role for complex cells (Sasaki et al., 2010). Third, it leaves out different ways to get the position and orientation information, such as eye fixations (Intoy et al., 2021). And fourth, it avoids the delicate question of whether the max operation over a column (Equation 12) truly captures a tangent element. Nevertheless, since our focus is geometric, it does capture all of the necessary ingredients and simplifies computations.

This binocular model allows us to define disparity and frontoparallel coordinates as

$$\begin{cases} d = \frac{x_L - x_R}{2} \\ x = \frac{x_R + x_L}{2}, \end{cases} \tag{16}$$

perfectly in accordance with the introduction of cyclopean coordinates in (4). In this way (x, y, d) correspond to the neural correlate of (r_1, r_2, r_3), via the change of variables (5).

3.2 The cortical fiber bundle of binocular cells

The hypercolumnar structure of monocular simple cells (orientation selective) has been described as a jet fiber bundle in the works of Petitot and Tondut (1999), among many others. We concentrate on the fiber bundle $\mathbb{R}^2 \times \mathbb{S}^1$, with fiber $\mathbb{S}^1$; see, e.g., Ben-Shahar and Zucker, 2004 among many others.

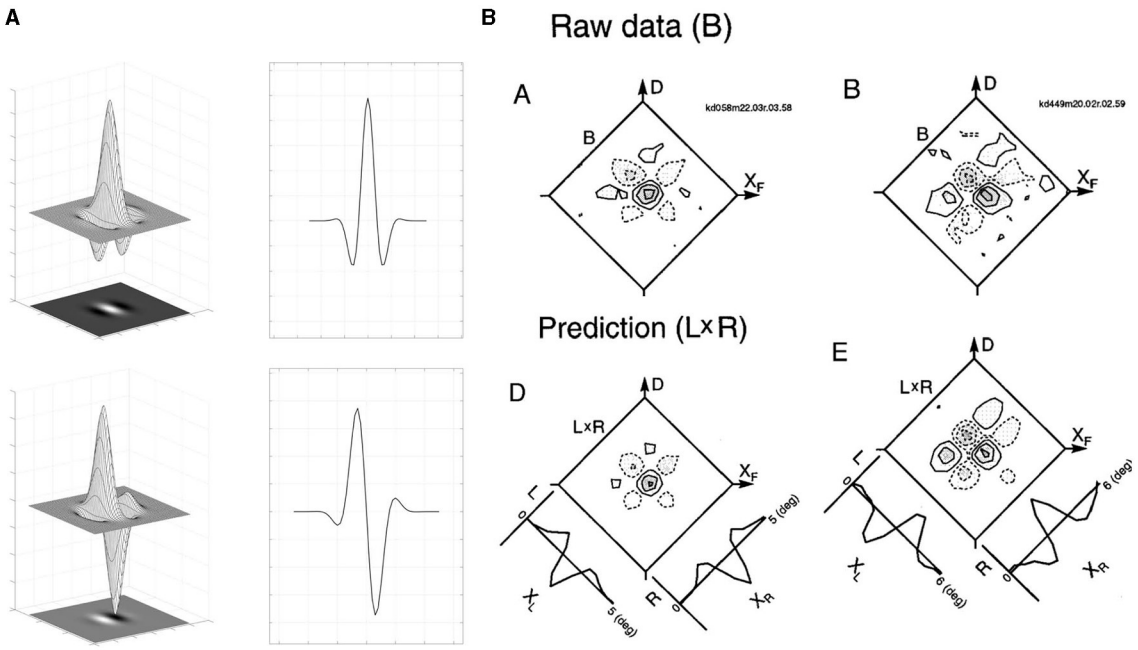


FIGURE 7
(A) Even (top) and odd (bottom) part of Gabor function: the surface of the two-dimensional filters, their common bi-dimensional representation and a mono-dimensional section. **(B)** Comparisons between binocular interaction RPs and the product of left and right eye RPs, where left and right RPs are shown in image **(A)**. Binocular interaction RPs (Raw data) of a cell is shown on the top row. Contour plots for the product of left and right eye RPs (L × R) are shown in the bottom row along with 1-dimensional profiles of the left (L) and right (R) eye RPs. Figure adapted from Anzai et al. (1999b).

In our setting, the binocular structure is based on monocular ones; recall the example illustrations from the Introduction. In particular, for each cell on the left eye there is an entire fiber of cells on the right, and vice versa, for each cell on the right there is an entire fiber of cells on the left. This implies that the binocular space is equipped with a symmetry that involves the left and right structures, allowing us to use the cyclopean coordinates (x, y, d) defined in (16).

Hence, we define the cyclopean retina $\mathcal{R}$, identified with $\mathbb{R}^2$, endowed with coordinates (x, y). The structure of the fiber is $\mathcal{F} = \mathbb{R} \times \mathbb{S}^1 \times \mathbb{S}^1$, with coordinates $(d, \theta_L, \theta_R) \in \mathcal{F}$. The total space is defined in a trivial way, $\mathcal{E} = \mathcal{R} \times \mathcal{F} = \mathbb{R}^2 \times \mathbb{R} \times \mathbb{S}^1 \times \mathbb{S}^1$, and the projection $\pi : \mathcal{E} \longrightarrow \mathcal{R}$ is the trivial projection $\pi(x, y, d, \theta_L, \theta_R) = (x, y)$. The preimage of the projection $\mathcal{E}_{(x,y)} := \pi^{-1}(\{(x, y)\})$, for every $(x, y) \in \mathcal{R}$, is isomorphic to the fiber $\mathcal{F}$, and the local trivialization property is naturally satisfied.

A schematic representation can be found in Figure 8. The base has been depicted as 1-dimensional, considering the restriction $\mathcal{R}_{|x}$ of the cyclopean retina $\mathcal{R}$ on the coordinate x. The left image displays only the disparity component of the fiber $\mathcal{F}$, encoding the relationships between left and right retinal coordinates. The right image shows the presence of the left and right monodimensional orientational fibers.

3.3 Binocular energy model

To simplify calculations, as stated in the Introduction, we follow the classical binocular energy model (Anzai et al., 1999b) for binocular RPs. The basic idea is a binocular neuron receives input from each eye; if the sum $O_L + O_R$ of the inputs from the left and right eye is positive, the firing rate of the binocular neuron is proportional to the square of the sum, and it vanishes, if the sum of the inputs is negative:

$$O_B = (Pos(O_L + O_R))^2, \tag{17}$$

with $Pos(x) = \max\{x, 0\}$, O_B the binocular output.

If $O_L + O_R > 0$, then the output of the binocular simple cell can be explicitly written as $O_B = O_L^2 + O_R^2 + 2O_L O_R$. The first two terms represent responses due to monocular stimulation while the third term $2O_L O_R$ can be interpreted as the binocular interaction term.

The activity of a cell is then measured from the output and will be strongest at points that have a higher probability of matching each other. The maximum value over d of this quantity is the extracted disparity.

It is worth noting that neurophysiological computations of binocular profiles displayed in Figure 7B assume the mono-dimensionality of the monocular receptive profile, ignoring information about orientation of monocular simple cells. However, this information will be needed to encode different types of orientation disparity.

Remark 3.1 (Orientation matters). In 2001, the authors of Bridge et al. (2001) conducted investigations on the response of binocular neurons to orientation disparity, by extending the energy model of Anzai, Ohzawa, and Freeman to incorporate binocular differences in receptive-field orientation. More recently, the difference between

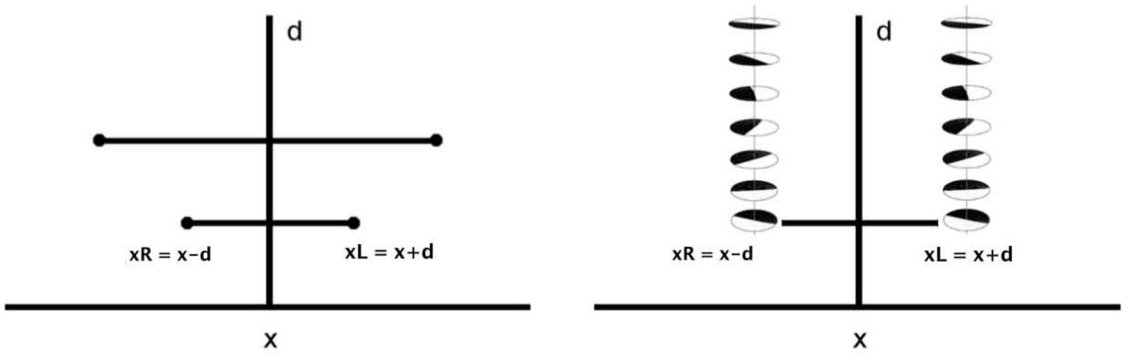


FIGURE 8
Left: schematic representation of the fiber bundle in two dimension, with relationships between left and right retinal coordinates. **Right**: representation of the selection of a whole fiber of left and right simple cells, for every x and for every d.

orientations in the receptive fields of the eyes has been confirmed (Sasaki et al., 2010).

The binocular energy model is a type of minimal model. It serves as a starting point, allowing the combination of monocular inputs. But is not sufficient to solve the stereo-matching problem.

Remark 3.2 (Connections). It is argued in Samonds et al. (2013) and Parker et al. (2016) that, in addition to the neural mechanisms that couple characteristics (such as signals, stimuli, or particular features) relating the left and right monocular structures, there must be a system of connections between binocular cells, which characterizes the processing mechanism of stereo vision; see also Samonds et al. (2013) in particular.

3.4 Differential interpretation of binocular RPs

It is possible to write the interaction term $O_L O_R$ coming from (17), in terms of the left and right receptive profiles:

$$\begin{aligned} O_L O_R = &\int \varphi_{\theta_L,x_L,y}(\tilde{x}_L,\tilde{y}_L) I_L(\tilde{x}_L,\tilde{y}_L) d\tilde{x}_L d\tilde{y}_L \\ &\int \varphi_{\theta_R,x_R,y}(\tilde{x}_R,\tilde{y}_R) I_R(\tilde{x}_R,\tilde{y}_R) d\tilde{x}_R d\tilde{y}_R \\ = &\int\int \varphi_{\theta_L,x_L,y}(\tilde{x}_L,\tilde{y}_L)\varphi_{\theta_R,x_R,y}(\tilde{x}_R,\tilde{y}_R) I_L(\tilde{x}_L,\tilde{y}_L) \\ &I_R(\tilde{x}_R,\tilde{y}_R) d\tilde{x}_R d\tilde{y}_R d\tilde{x}_L d\tilde{y}_L. \end{aligned} \tag{18}$$

If we fix $(\tilde{x}_R,\tilde{y}_R,\tilde{x}_L,\tilde{y}_L)$, we derive the expression of the binocular profiles $\varphi_{L,R} = \varphi_{\theta_R,x_R,y}\varphi_{\theta_L,x_L,y}$ as the product of monocular left and right profiles. This is in accordance with the measured profiles of Figure 7B).

Proposition 3.1. The binocular interaction term can be associated with the cross product of the left and right directions defined through (13), namely $\omega_L^\star$ and $\omega_R^\star$ of monocular simple cells:

$$O_L O_R = \omega_L^\star \times \omega_R^\star. \tag{19}$$

Proof. The idea is that the binocular output is the combined result of the left and right actions of monocular cells, thus identifying a direction in the space of cyclopean coordinates. The detailed proof of this proposition can be found in Appendix B (Supplementary material). □

To better understand the geometrical idea behind Proposition 3.1, we recall that the retinal coordinates can be expressed in terms of cyclopean coordinates (4) as $x_R = x - d$ and $x_L = x + d$, and so we can write $\omega_L^\star$ and $\omega_R^\star$ in the $3D$ space of coordinates (x, y, d) as:

$$\begin{aligned} \omega_R^\star &= (-\sin\theta_R, \cos\theta_R, \sin\theta_R)^T \\ \omega_L^\star &= (-\sin\theta_L, \cos\theta_L, -\sin\theta_L)^T. \end{aligned} \tag{20}$$

We define $\omega_{bin} := \omega_L^\star \times \omega_R^\star$ as the natural direction characterizing the binocular structure:

$$\omega_{bin} = \begin{pmatrix} \sin(\theta_R + \theta_L) \\ 2\sin\theta_R \sin\theta_L \\ \sin(\theta_R - \theta_L) \end{pmatrix}. \tag{21}$$

Remark 3.3. The vector ω_{bin} of Equation (21) can be interpreted as the intersection of the orthogonal spaces defined with respect to $\omega_R^\star$ and $\omega_L^\star$ when expressed in cyclopean coordinates (x, y, d). More precisely, if

$$\begin{aligned} (\omega_L^\star)^\perp &= \text{span}\left\{ \begin{pmatrix} \cos\theta_L \\ \sin\theta_L \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \right\}, \\ (\omega_R^\star)^\perp &= \text{span}\left\{ \begin{pmatrix} \cos\theta_R \\ \sin\theta_R \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \right\} \end{aligned} \tag{22}$$

then

$$\omega_{bin} = (\omega_L^\star)^\perp \cap (\omega_R^\star)^\perp. \tag{23}$$

The result of the intersection of these monocular structures identifies a direction, as shown in Figure 9A.

We earlier showed that the result of the action of a monocular odd simple cell is to select directions for the propagation of information. We now combine these, for the two eyes, to show that in the three-dimensional case the binocular neural mechanisms also lead to a direction. We will see in the next sections that this direction is the direction of the tangent vector to the 3*D* stimulus, provided points are corresponding.

3.5 Compatibility with stereo geometry

We consider the direction characterizing the binocular structure ω_{bin} defined in (21) and we show that it can be associated with the 3*D* tangent vector to the 3*D* curve. The idea is that this tangent vector is orthogonal both to $\omega_R^\star$ and to $\omega_L^\star$, and therefore it has the direction of the vector product $\omega_L^\star \times \omega_R^\star$.

Precisely, we consider the normalized tangent vector t_L and t_R on retinal planes

$$t_R = (\cos\theta_R, \sin\theta_R)^T \quad t_L = (\cos\theta_L, \sin\theta_L)^T, \tag{24}$$

to the points (x_R, y) and (x_L, y) respectively. Taking into account that f is the focal coordinate of the retinal planes in $\mathbb{R}^3$, then we associate to these points the correspondents in $\mathbb{R}^3$, namely $\tilde{m}_L = (x_L - c, y, f)^T$, $\tilde{m}_R = (x_R + c, y, f)^T$. Applying Equation (7), it is possible to derive the tangent vector of the three dimensional contour:

$$\begin{aligned} U_{t_L} &= P_L^{-1}\tilde{m}_L \times P_L^{-1}\tilde{t}_L = \begin{pmatrix} x_L \\ y_L \\ f \end{pmatrix} \times \begin{pmatrix} \cos\theta_L \\ \sin\theta_L \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} -f\sin\theta_L \\ f\cos\theta_L \\ x_L\sin\theta_L - y_L\cos\theta_L \end{pmatrix}, \\ U_{t_R} &= P_R^{-1}\tilde{m}_R \times P_R^{-1}\tilde{t}_R = \begin{pmatrix} x_R \\ y_R \\ f \end{pmatrix} \times \begin{pmatrix} \cos\theta_R \\ \sin\theta_R \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} -f\sin\theta_R \\ f\cos\theta_R \\ x_R\sin\theta_R - y_R\cos\theta_R \end{pmatrix}, \end{aligned} \tag{25}$$

and the tangent direction is recovered by

$$U_{t_L} \times U_{t_R} = f \begin{pmatrix} \frac{x_L+x_R}{2}\sin(\theta_R - \theta_L) - \frac{x_R-x_L}{2}\sin(\theta_L + \theta_R) \\ y\sin(\theta_R - \theta_L) - (x_R - x_L)(\cos(\theta_R - \theta_L) - \cos(\theta_L + \theta_R)) \\ f\sin(\theta_R - \theta_L) \end{pmatrix} \tag{26}$$

If we define

$$\tilde{\omega}_L^\star := \frac{d}{fc}U_{t_L}, \qquad \tilde{\omega}_R^\star := \frac{d}{fc}U_{t_R} \tag{27}$$

and the corresponding 2 form $\omega_{\mathbb{R}^3} := \tilde{\omega}_L^\star \times \tilde{\omega}_R^\star$, using the change of variables (16) we observe that:

$$\tilde{\omega}_L^\star = \omega_L^\star, \quad \tilde{\omega}_R^\star = \omega_R^\star, \quad \omega_{\mathbb{R}^3} = \omega_{bin}, \tag{28}$$

up to a scalar factor. See Appendix C (Supplementary material) for explicit computation.

In this way, the disparity binocular cells couple in a natural way positions, identified with points in $\mathbb{R}^3$, and orientations in $\mathbb{S}^2$, identified with three-dimensional unitary tangent vectors. As already observed in Remark 3.2, the geometry of the stereo vision is not solved only with these punctual and directional arguments, but there is the need to take into accounts suitable type of connections. In Alibhai and Zucker (2000); Li and Zucker (2003, 2006), Zucker et al. proposed a model that considered the curvature of monocular structures as an additional variable. Instead, we propose to consider simple monocular cells selective for orientation, and to insert the notion of curvature directly into the definition of connection. It is therefore natural to introduce the perceptual space via the manifold $\mathbb{R}^3 \rtimes \mathbb{S}^2$, in line with the theoretical toolbox proposed in Miolane and Pennec (2016) to generalize 2D neurogeometry to 3D images, and adapt this framework to our problem, looking for appropriate curves.

3.6 A perceptual model in the space of 3D position-orientation

We now derive the objects in Figure 3A. We have clarified (end of Section 3.5) that binocular cells are parameterized by points in $\mathbb{R}^3$, and orientations in $\mathbb{S}^2$. An element ξ of the space $\mathbb{R}^3 \rtimes \mathbb{S}^2$ it is defined by a point $p = (p1, p2, p3)$ in $\mathbb{R}^3$ and an unitary vector $n \in \mathbb{S}^2$. Since the topological dimension of this geometric object is 2, we introduce the classical spherical coordinates (θ, φ) such that $n = (n_1, n_2, n_3) \in \mathbb{S}^2$ can be parameterized as:

$$\begin{aligned} n_1 &= \cos\theta\sin\varphi \\ n_2 &= \sin\theta\sin\varphi \\ n_3 &= \cos\varphi \end{aligned} \tag{29}$$

with $\theta \in [0, 2\pi]$ and $\varphi \in (0, \pi)$. The ambiguity that arises using local coordinate chart is overcome by the introduction of a second chart, covering the singular points.

Translations and rotations are expressed using the group law of the three-dimensional special Euclidean group $SE(3)$, defining the group action

$$\sigma : \mathbb{R}^3 \rtimes \mathbb{S}^2 \times SE(3) \longrightarrow \mathbb{R}^3 \rtimes \mathbb{S}^2 \text{ s.t. } \sigma((p,n),(q,R)) = (Rp+q, Rn), \tag{30}$$

with $(p, n) \in \mathbb{R}^3 \rtimes \mathbb{S}^2$, $(q, R) \in SE(3)$, namely $R \in SO(3)$ tridimensional rotation, and $q \in \mathbb{R}^3$.

3.6.1 Stereo sub-Riemannian geometry

The emergence of a privileged direction in $\mathbb{R}^3$ (associated with the tangent vector to the stimulus) is the reason why we endow $\mathbb{R}^3 \rtimes \mathbb{S}^2$ with a sub-Riemannian structure that favors the direction in 3*D* identified by ω_{bin}.

Formally, we consider admissible movements in $\mathbb{R}^3 \rtimes \mathbb{S}^2$ described by vector fields:

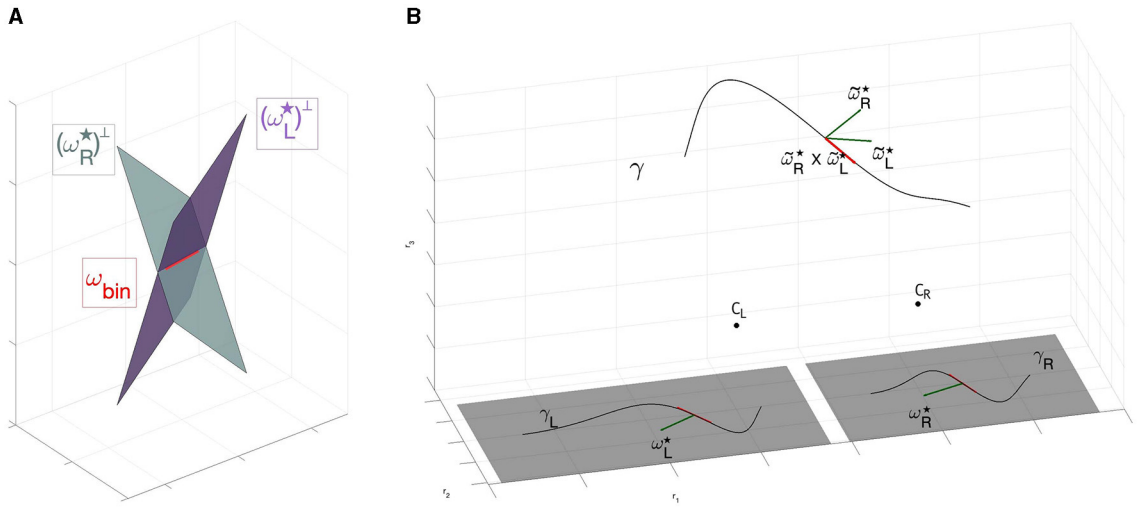


FIGURE 9
(A) Direction detected by ω_{bin} through the intersection of left and right planes generated by $(\omega_R^\star)^\perp$ and $(\omega_L^\star)^\perp$. Red vector corresponds to the associated 2-form ω_{bin}. **(B)** Three dimensional reconstruction of the space from retinal planes. The 1- forms $\omega_L^\star$ and $\omega_R^\star$ are identified with the normal to the curves γ_L and γ_R. Their three dimensional counterpart $\tilde{\omega}_L^\star$ and $\tilde{\omega}_R^\star$ identify the tangent vector to the curve $\gamma : \mathbb{R} \to \mathbb{R}^3$ by the cross product $\tilde{\omega}_L^\star \times \tilde{\omega}_R^\star$.

$$\begin{aligned} Y_{\mathbb{R}^3,\xi} &= \sin\varphi\cos\theta\,\partial_1 + \sin\varphi\sin\theta\,\partial_2 + \cos\varphi\,\partial_3 \\ Y_{\theta,\xi} &= -\frac{1}{\sin\varphi}\,\partial_\theta \\ Y_{\varphi,\xi} &= \partial_\varphi \end{aligned} \tag{31}$$

with $\xi \in \mathbb{R}^3 \rtimes \mathbb{S}^2$ for $\varphi \neq 0, \varphi \neq \pi$. The admissible tangent space[9] at a point ξ

$$\mathcal{A}_\xi := \mathrm{span}\{Y_{\mathbb{R}^3,\xi}, Y_{\theta,\xi}, Y_{\varphi,\xi}\} \tag{32}$$

encodes the coupling between position and orientations, as remarked by Duits and Franken (2011). In particular, the vector field $Y_{\mathbb{R}^3}$ identifies the privileged direction in $\mathbb{R}^3$, while Y_θ and Y_φ allow changing this direction, involving just orientation variables of $\mathbb{S}^2$. The vector fields $\{Y_{\mathbb{R}^3}, Y_\theta, Y_\varphi\}$ and their commutators generate the tangent space of $\mathbb{R}^3 \rtimes \mathbb{S}^2$ in a point, allowing to connect every point of the manifold using privileged directions (*Hörmander condition*). Furthermore, it is possible to define a sub-Riemannian structure by choosing a scalar product on the admissible tangent bundle $\mathcal{A}$: the simplest choice is to declare the vector fields $\{Y_{\mathbb{R}^3}, Y_\theta, Y_\varphi\}$ orthonormal, considering on $\mathbb{S}^2$ the distance inherited from the immersion in $\mathbb{R}^3$ with the Euclidean metric.

3.6.2 Change of variables

We have already expressed the change of variable in the variables (x, y, d) to (r_1, r_2, r_3) in Equation (5). However, the cortical coordinates also contain the angular variables θ_R and θ_L which involve the introduction of the spherical coordinates θ, φ.

9 see Appendix A (Supplementary material) for the definition of admissible tangent space.

To identify a change of variable among these variables, we first introduce the function $(r_1, r_2, r_3, \theta, \varphi) \xrightarrow{F} (x, y, d, \theta_L, \theta_R)$:

$$F: \quad \mathbb{R}^3 \rtimes \mathbb{S}^2 \longrightarrow \mathbb{R}^3 \rtimes \mathbb{S}^2$$
$$\begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ \theta \\ \varphi \end{pmatrix} \mapsto \begin{pmatrix} \frac{fr_1}{r_3} \\ \frac{fr_2}{r_3} \\ \frac{cf}{r_3} \\ \tan^{-1}\left(\frac{r_3\sin\theta\cos\varphi - r_2\cos\varphi}{r_3\cos\theta\sin\varphi - (c+r_1)\cos\varphi}\right) \\ tan^{-1}\left(\frac{r_3\sin\theta\cos\varphi - r_2\cos\varphi}{r_3\cos\theta\sin\varphi - (c-r_1)\cos\varphi}\right) \end{pmatrix}, \tag{33}$$

where the retinal right angle $\theta_R = \tan^{-1}\left(\frac{r_3\sin\theta\cos\varphi - r_2\cos\varphi}{r_3\cos\theta\sin\varphi - (c+r_1)\cos\varphi}\right)$ and the retinal left angle $\theta_L = \tan^{-1}\left(\frac{r_3\sin\theta\cos\varphi - r_2\cos\varphi}{r_3\cos\theta\sin\varphi - (c-r_1)\cos\varphi}\right)$ are obtained considering Equation (6).

Analogously, it is possible to define the change of variable $(x, y, d, \theta_L, \theta_R) \xrightarrow{G} (r_1, r_2, r_3, \theta, \varphi)$:

$$G: \quad \mathbb{R}^3 \rtimes \mathbb{S}^2 \longrightarrow \mathbb{R}^3 \rtimes \mathbb{S}^2$$
$$\begin{pmatrix} x \\ y \\ d \\ \theta_R \\ \theta_L \end{pmatrix} \mapsto \begin{pmatrix} \frac{cx}{d} \\ \frac{cy}{d} \\ \frac{cf}{d} \\ \tan^{-1}\left(\frac{2\sin\theta_R\sin\theta_L}{\sin(\theta_R+\theta_L)}\right) \\ \tan^{-1}\left(\frac{\sqrt{\sin^2(\theta_R+\theta_L)+4\sin^2\theta_R\sin^2\theta_L}}{\sin(\theta_R-\theta_L)}\right) \end{pmatrix}, \tag{34}$$

where the angles $\theta = \tan^{-1}\left(\frac{2\sin\theta_R\sin\theta_L}{\sin(\theta_R+\theta_L)}\right)$ and $\varphi = \tan^{-1}\left(\frac{\sqrt{\sin^2(\theta_R+\theta_L)+4\sin^2\theta_R\sin^2\theta_L}}{\sin(\theta_R-\theta_L)}\right)$ are obtained considering that $\tan\theta = \frac{(\vec{Y}_{\mathbb{R}^3})_2}{(\vec{Y}_{\mathbb{R}^3})_1}$ and $\tan\varphi = \frac{\sqrt{(\vec{Y}_{\mathbb{R}^3})_1^2 + (\vec{Y}_{\mathbb{R}^3})_2^2}}{(\vec{Y}_{\mathbb{R}^3})_3}$.

3.6.3 Integral curves

The connectivity of the space is described by admissible curves of the vector fields spanning $\mathcal{A}$. In particular, a curve $\Gamma : [0, T] \longrightarrow \mathbb{R}^3 \rtimes \mathbb{S}^2$ is said to be *admissible*[10] if:

$$\dot{\Gamma}(t) \in \mathcal{A}_{\Gamma(t)}, \leftrightarrow \dot{\Gamma}(t) = a(t)\vec{Y}_{\mathbb{R}^3,\Gamma(t)} + b(t)\vec{Y}_{\theta,\Gamma(t)} + c(t)\vec{Y}_{\varphi,\Gamma(t)}, \tag{35}$$

where a, b, c are sufficiently smooth function on $[0, T]$. We will consider a particular case of these admissible curves, namely constant coefficient integral curves with $a(t) = 1$, since the vector field $Y_{\mathbb{R}^3}$ represents the tangent direction of the $3D$ stimulus (and so it never vanishes):

$$\dot{\Gamma}(t) = \vec{Y}_{\mathbb{R}^3,\Gamma(t)} + c_1\vec{Y}_{\theta,\Gamma(t)} + c_2\vec{Y}_{\varphi,\Gamma(t)}, \tag{36}$$

with c_1 and c_2 varying in $\mathbb{R}$.

These curves can be thought of in terms of trajectories in $\mathbb{R}^3$ describing a movement in the $\vec{Y}_{\mathbb{R}^3}$ direction, which can eventually change according to $\vec{Y}_\theta$ and $\vec{Y}_\varphi$. An example of the fan of integral curves was shown in the Introduction in Figure 3B.

It is worth noting that in the case described by coefficients c_1 and c_2 equal to zero, the $3D$ trajectories would be straight lines in $\mathbb{R}^3$; by varying the coefficients c_1 and c_2 in $\mathbb{R}$, we allow the integral curves to follow curved trajectories, twisting and bending in all space directions.

Formally, the amount of "twisting and bending" in space is measured by introducing the notions of curvature and torsion. We then investigate how these measurements are encoded in the parameters of the family of integral curves, and what constraints have to be imposed to obtain different typologies of curves.

Remark 3.4. The $3D$ projection of the integral curves (36) will be denoted γ and satisfy $\dot{\gamma}(t) = (\cos\theta(t)\sin\varphi(t), \sin\theta(t)\sin\varphi(t), \cos\varphi(t))^T$. Classical instruments of differential geometry let us compute the curvature and the torsion of the curve $\gamma(t)$:

$$\begin{aligned} k &= \sqrt{(\dot{\varphi})^2 + \sin^2\theta(\dot{\theta})^2}, \\ \tau &= \frac{1}{k^2}(-\cos\varphi\sin^2\varphi(\dot{\theta})^3 - \sin\varphi\dot{\varphi}\ddot{\theta} + \dot{\theta}(-2\cos\varphi(\dot{\varphi})^2 + \sin\varphi\ddot{\varphi})). \end{aligned} \tag{37}$$

Using the explicit expression of the vector fields Y_θ and Y_φ in Equation (36), we get

$$\dot{\theta} = -\frac{c_1}{\sin\varphi}, \qquad \dot{\varphi} = c_2, \tag{38}$$

from which it follows that:

$$\begin{aligned} k &= \sqrt{c_1^2 + c_2^2} \\ \tau &= c_1 \operatorname{cotan}\varphi. \end{aligned} \tag{39}$$

Proposition 3.2. By varying the parameters c_1 and c_2 in (39) where we explicitly find solutions of (36), we have:

1. If $\varphi = \frac{\pi}{2}$ then $k = \sqrt{c_1^2}, \tau = 0$, and so the family of curves (36) are circles of radius $1/c_1^2$ on the fronto-parallel plane $r_3 = \text{cost}$.
2. If $\varphi = \varphi_0$, with $\varphi_0 \neq \pi/2$, then $k = \sqrt{c_1^2}$ and $\tau = c_1 \operatorname{cotan}\varphi_0$, and so the family of curves (36) are r_3-helices.
3. If $\theta = \theta_0$ then $k = \sqrt{c_2^2}$, $\tau = 0$, and so the family of curves (36) are circles of radius $1/c_2^2$ in the osculating planes.

Proof. The computation follows immediately from the computed curvature and torsion of (39) and classical results of differential geometry. □

Remark 3.5. If we know the value of the curvature k, and we have one free parameter, c_2, in the definition of the integral curves (36), then we are in the setting of Proposition 2.1. In fact, the coefficient c_1 is obtained by imposing $c_1 = \pm\sqrt{k^2 - c_2^2}$, and in particular the component that remains to be determined is the torsion.

Examples of particular cases of the integral curves (36) according to Proposition 3.2 and Remark 3.5 are visualized in Figure 10.

4 Comparison with experimental data

Our sub-Riemannian model enjoys some consistency with the biological and psychophysical literature. We here describe some initial connections.

4.1 Biological connections

The foundation for building our sub-Riemannian model of stereo was a model of curve continuation, motivated by the orientation columns at each position. The connections between cells in nearby columns were, in turn, a geometric model of long-range horizontal connections in visual cortex (Bosking et al., 1997). In the Introduction we cartooned aspects of the cortical architecture that support binocular processing. Although the inputs are organized into ocular dominance bands, there is no direct evidence for "stereo columns" in V1 analogous to the monocular orientation columns. But such columns are not strictly necessary for our model. Rather, what is central is how information propagates. We showed in Figure 1C that there is evidence of long-range connections between binocular cells, and our model informs, abstractly, what information could propagate along these connections. Although less extensive than in the monocular case, some measurements are beginning to emerge that are informative.

The Grinwald group first established the presence of long-range connections between binocular cells (Malach et al., 1993) (see also Figure 11A), using biocytin. This is a molecule that is taken up by neurons, propagates directly along neuronal processes and is deposited at excitatory synapses. These results were refined, more recently, by the Fitzpatrick group (Scholl et al., 2022), using *in vivo* calcium imaging. As shown in Figure 1C the authors demonstrated both the monocular and the binocular inputs for stereo, and (not shown) the dependence on orientations.

More precisely, Malach et al. (1993) showed selective *anisotripic* connectivity among binocular regions: the biocytin tracer does not spread uniformly, but rather is highly directional with

10 sometimes the term *horizontal* is preferred.

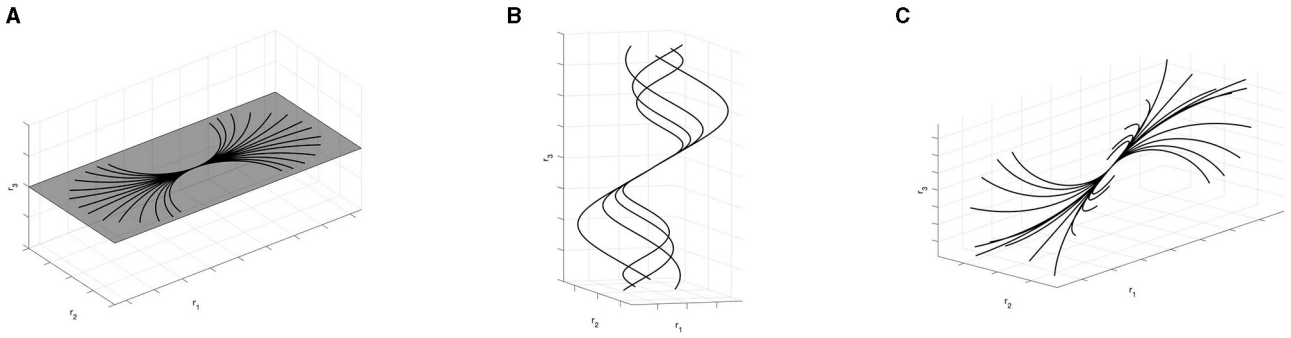


FIGURE 10
Examples of integral curves obtained varying parameters c_1 and c_2. **(A)** Arc of circles for $\varphi = \pi/2$. **(B)** r_3-helices for $\varphi = \pi/3$. **(C)** Family of curves with constant curvature k and varying torsion parameter.

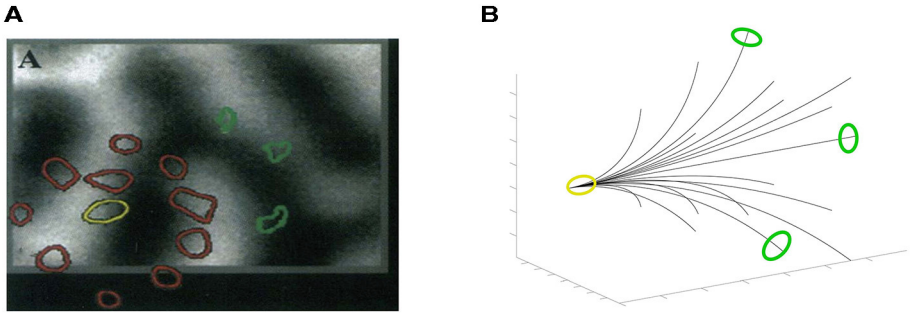


FIGURE 11
(A) A biocytin injection superimposed on a map of ocular dominance columns, image result from the work in Malach et al. (1993). Binocular zones are in the middle of monocular zones (coded in black and white). Starting from the injection site (yellow circle in the center of a binocular zone) the patches' propagation (red corresponds to dense while green to sparsely labeled) tends to avoid highly monocular sites, bypassing the centers of ocular dominance columns, and are located in binocular zones. **(B)** 3*D* interpretation of the physiological image **(A)**.

distance from the injection point. (This was the case with monocular biocytin injections as well.) Putting this together with Scholl et al. (2022), we interpret the anisotropy as being related to (binocular) orientation (Scholl et al., 2022), which is what the integral curves of our vector fields suggest. Our 3*D* association fields are strongly directional, and information propagates preferentially in the direction of (the starting point of) the curve. An example can be seen in Figure 11B, where the fan of integral curves (36) is represented, superimposed with colored patches, following the experiment proposed in Malach et al. (1993). We look forward to more detailed experiments along these lines.

4.2 Psychophysics and association fields

In this section, we show that the connections described by the integral curves in our model can be related to the geometric relationships from psychophysical experiments on perceptual organization of oriented elements in $\mathbb{R}^3$; in other words, that our connections serve as a generalization of the concept of an association field in 3*D*.

4.2.1 Toward a notion of *association field* for 3D contours

The perception of continuity between two elements of position-orientation in $\mathbb{R}^3$ has been studied experimentally. To start, Kellman, Garrigan, and Shipley (Kellman et al., 2005a,b) introduce *3D relatability*, as a way to extend to 3*D* the experiments of Field, Hayes and Hess (Field et al., 1993) in 2*D*.

Particularly, in a system of 3*D* Cartesian coordinates, it is possible to introduce oriented edges E at the application point $(r_1, r_2, r_3)^T$ and with an orientation identified with the angles θ and φ. This orientation can be read, in our case, through the direction expressed by $(\cos\theta \sin\varphi, \sin\theta \sin\varphi, \cos\varphi)^T$. For an initial edge E_0, with application point on the origin of the coordinate system $(0,0,0)^T$ and orientation lying on the r_1-axis, described by $\theta = 0, \varphi = \pi/2$, the range of possible orientations (θ, φ)[11] for 3*D*-relatable edges with E_0 is given by:

11 The angle φ here has been modified to be compatible with our set of coordinates. The relationship between the angle $\tilde{\varphi}$ in works (Kellman et al., 2005a,b) can be expressed as : $\tilde{\varphi} = \text{acos}(\sin\varphi) + \pi$.

$$\tan^{-1}\left(\frac{r_2}{r_1}\right) \leq \theta \leq \frac{\pi}{2} \quad \text{and} \quad \frac{\pi}{2} \leq \frac{3\pi}{2} - \varphi \leq \tan^{-1}\left(\frac{r_3}{r_1}\right). \tag{40}$$

The bound on these equations identified with the quantity $\frac{\pi}{2}$ incorporates the 90 degree constraint in three dimensions, while the bounds defined by the inverse of the tangent express the absolute orientation difference between the reference edge E_0 and an edge positioned at the arbitrary oriented point $E_{(r_1,r_2,r_3)}$ so that its linear extension intersects E_0; see Kellman et al. (2005a,b) for further details.

Numerical simulations allow us to visually represent an example of the 3*D* positions and orientations that meet the 3*D* relatability criteria. Starting from an initial edge E_0 with endpoints in $(p_{01}, p_{02}, p_{03})^T$ and orientation on the e_1- axis, we represent for an arbitrary point $(p_1, p_2, p_3)^T$ the limit of the relatable orientation (θ, φ). Results are shown in Figure 12A.

Remark 4.1. By projecting on the retinal planes of the 3*D* fan of relatable points, it is possible to notice that these projections are in accordance with the notion of 3*D* compatibility field in Alibhai and Zucker (2000) (see Figures 12B, C).

Psychophysical studies, see Hess and Field (1995); Hess et al. (1997); Deas and Wilcox (2015), have investigated the properties of the curves that are suitable for connecting these relatable points. These curves are well-described as smooth and monotonic. In particular, using non-oriented contour elements for contours, Hess et al. (1997) indicate that contour elements can be effectively grouped based primarily on the good continuation of contour elements in depth. This statement is confirmed by the more recent work of Deas and Wilcox (2015) who, in addition, observe that detection of contours defined by regular depth continuity is faster than detection of discontinuous contours. All these results support the existence of depth grouping operations, arguing for the extension of Gestalt principles of continuity and smoothness in three dimensional space. Finally, on the relationship of the three-dimensional curves to 2-dimensional association fields, see Kellman et al. (2005b); Khuu et al. (2016). These authors have assumed that the strength of the relatable edges in the co-planar planes of E_0 must meet the relations of the bi-dimensional association fields of Field et al. (1993).

4.2.2 Compatibility with the sub-Riemannian model

To model the associations underlying the 3*D* perceptual organization discussed in the previous paragraph, we consider again the constant coefficient family of integral curves studied in (36):

$$\dot{\Gamma}(t) = \vec{Y}_{\mathbb{R}^3,\Gamma(t)} + c_1 \vec{Y}_{\theta,\Gamma(t)} + c_2 \vec{Y}_{\varphi,\Gamma(t)}, \text{ with } c_1, c_2 \in \mathbb{R}. \tag{41}$$

Importantly, these curves locally connect the association fan generated by the geometry of 3*D* relatability. In particular, Figure 13B shows the family of the horizontal curves connecting the initial point E_0 with 3*D* relatable edges (Figure 13A). These curves are computed using Matlab solver function `ode45`.

In analogy with the experiment of Field, Hayes, and Hess in Field et al. (1993), we choose to represent non-relatable edges to the left of the starting point E_0, while on the right are 3*D* relatable edges. So, filled lines of the integral curves indicate the correlation between the central horizontal element E_0 and the ones on its right, while dotted lines connect the starting point E_0 with elements not correlated with it, as represented on the left part of the image.

Restricting the curves on the neighborhood of co-planar planes with an arbitrary edge *E*, we have different cases. First, on the r_1-r_2 plane (fronto-parallel) and the r_1-r_3 plane we have arcs of circle, as proved with Proposition 3.2. Furthermore, for an arbitrary plane in $\mathbb{R}^3$ containing an edge *E*, we observe that the curves generating with fixed angle φ are helices, and locally they satisfy the bidimensional constraint in the plane. Examples can be found in Figures 13C–E. In particular, the curves displayed in Figures 13C, D are well in accordance with the curves of the Citti-Sarti model, depicted in Figure 5.

One final connection with the psychophysical literature concerns how depth discrimination thresholds increase exponentially with distance (Burge, 2020 and references therein). This is related to how the fan of integral curves "spreads out" with distance (Figures 11, 12), which is also exponential. These notions are developed more fully in Bolelli et al. (2023a).

4.3 Integration of contours and stereo correspondence problem

Although the goal of this paper is not to solve the stereo correspondence problem, we can show how our geometry is helpful in understanding how to match left and right points and features. These ideas are developed more fully in Bolelli (2023).

Inspired by Hess and Field (1995), we consider a path stimulus γ interpreted as a contour, embedded in a background of randomly oriented elements: left and right retinal visual stimuli are depicted in Figure 14A. We perform an initial, simplified lift of the retinal images to a set $\Omega \subset \mathbb{R}^3 \rtimes \mathbb{S}^2$. This set contains all the possible corresponding points, obtained by coupling left and right points which share the same y retinal coordinate, see Figure 14B. The set Ω contains false matches, namely points that do not belong to the original stimulus. It is the task of correspondence to eliminate these false matches.

We compute for every lifted point the binocular output O_B of Equation (17). This output can be seen as a probability measure that gives information on the correspondence of the pair of left and right points. We then simply evaluate which are the points with the highest probability of being in correspondence, applying a process of suppression of the non-maximal pairs over the fiber of disparity. In this way, noise points are removed (Figure 14C). We now directly exploit the Gestalt rule of good continuation by filtering out any couple of elements with high curvature. This qualitative rule could be quantitatively modeled by considering the statistics of distribution of curvature and torsion in natural 3D images (Geisler and Perry, 2009). The remaining noise elements are orthogonal to the directions of the elements of the curve that we

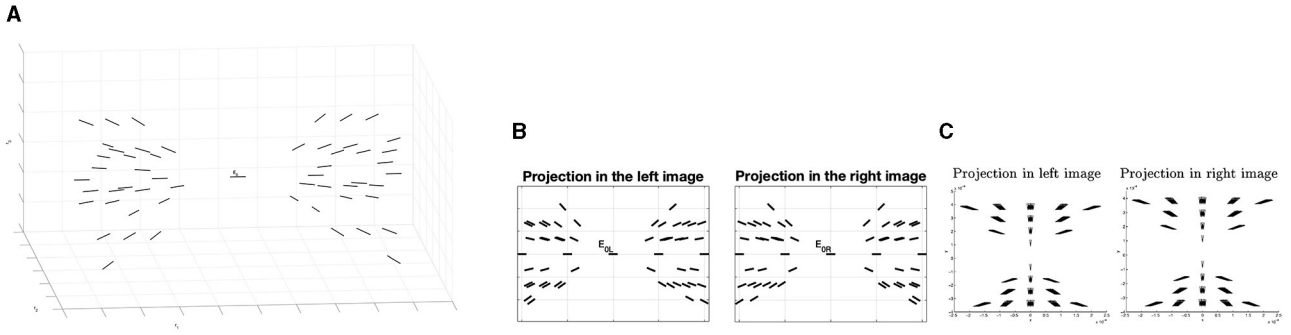


FIGURE 12
(A) Example of the fan of the 3*D* relatable edges with initial point E_0. **(B)** Example of 3*D* association field in the two left and right retinal planes, generated with the geometry of 3*D* relatability. **(C)** Example of 3*D* compatibility field of Alibhai and Zucker (2000).

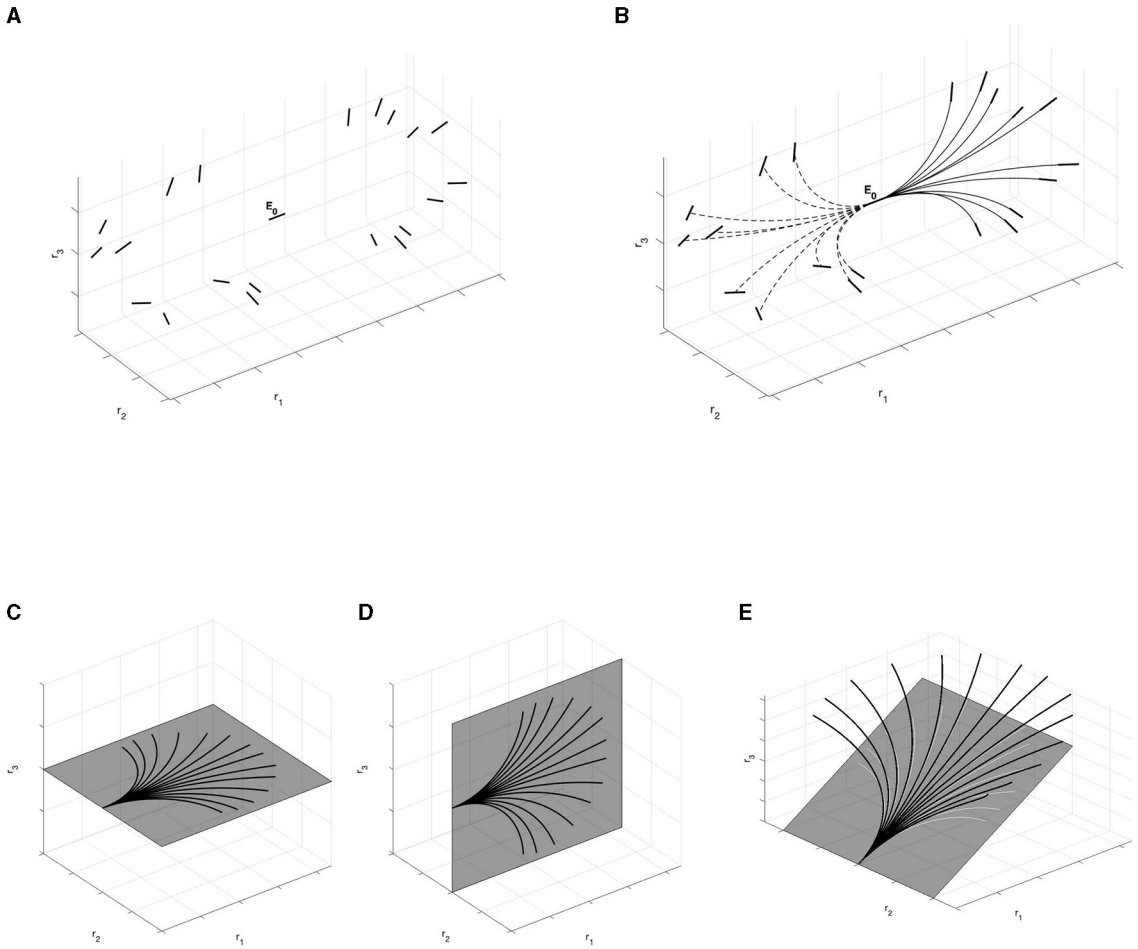


FIGURE 13
(A) 3*D* relatable edges displayed on the right of the initial edge E_0. Unrelatable 3*D* edges displayed on the left. **(B)** Horizontal integral curves with filled lines connect 3*D* relatable edges with initial point E_0. Horizontal integral curves with dotted lines do not connect 3*D* unrelatable edges. **(C)** Restriction of the fan of the integral curves on the e_1-e_2 plane. **(D)** Restriction of the fan of the integral curves on the e_1-e_3 plane. **(E)** Restriction of the fan at $\varphi = \varphi_0$. These curves (black lines) are not planar curves but helices. However, their projection (white lines) on the coplanar plane with initial edge satisfies the bidimensional constraints.

would like to reconstruct. Calculating numerically the coefficients c_1 and c_2 of integral curves (36) that connect all the remaining pairs of points, we can obtain for every pair the value of curvature and torsion using (39).

Figures 14E, F show matrices M representing the values of curvature or torsion for every pair of points ξ_i, ξ_j in the element M_{ij}. In particular, we observe that random points are characterized by very high curvature and deviating torsion. So, by discarding

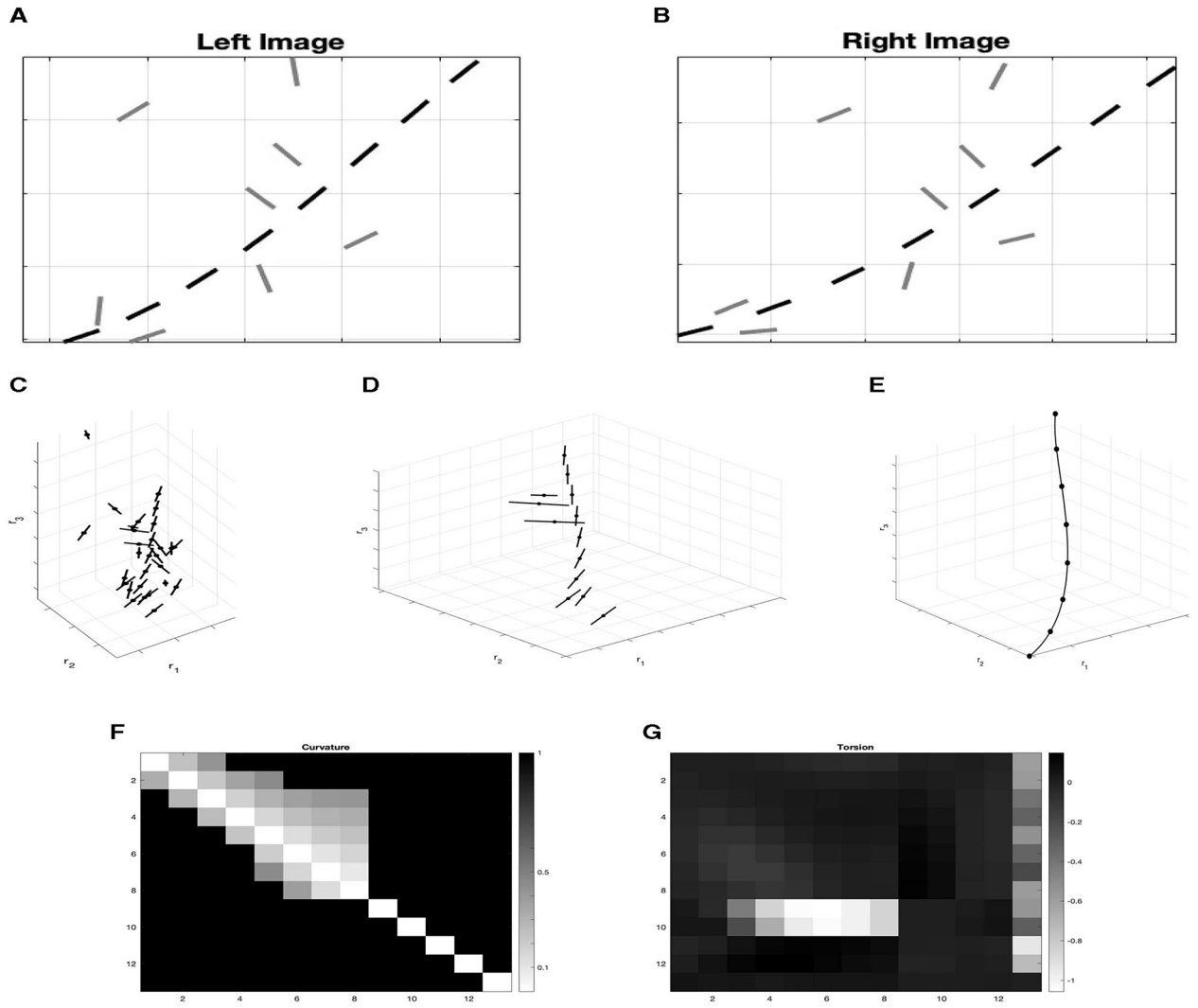


FIGURE 14
(A) Left and right retinal images of the set Ω. Black points are the projection of the point of the curve γ, while gray points are background random noise. **(B)** Lifting of the two left and right retinal images of image **(A)** in the space of position and orientation $\mathbb{R}^3 \times \mathbb{S}^2$. **(C)** Selection of lifted points according to the binocular output. **(D)** Points of the stimulus γ connected by integral curves (36). **(E, F)** Matrices M which element M_{ij} represents the value of curvature/ torsion for every couple of points ξ_i, ξ_j. The first eight points correspond to points of the curve γ while the others are random noise. **(E)** Curvature matrix. **(F)** Torsion matrix.

these high values, we select only the three-dimensional points of the curve γ, which are well-connected by the integral curves, as shown in Figure 14D. This is in accordance with the idea developed in Alibhai and Zucker (2000); Li and Zucker (2003, 2006), where curvature and torsion provide constraints for reconstruction in 3D.

In this artificial example we assumed that local edge elements have already been detected. Our goal was simply proof-of-concept. To apply this approach to realistic images, of course, stages of edge detection would have to be adopted, for which there is a huge literature well outside the scope of our theoretical study.

5 Summary and conclusions

Understanding good continuation in depth, like good continuation for planar contours, can benefit from basic physiological constraints; from psychophysical performance measures, and from mathematical modeling. In particular, good continuation in the plane is supported by orientation selectivity and cortical architecture (orientation columns), by association field grouping performance, and by geometric modeling. We showed that the same should be true for good continuation in depth. However, while the psychophysical data may be comparable, the physiological data are weaker and the geometry of continuation is not as well-understood. In this paper, we introduced the neuro-geometry of stereo vision to fill this gap. It is strongly motivated by an analogical extension to $3D$ of $2D$ geometry, while respecting the psychophysics. In the end, it allowed us to be precise about the type of geometry that is relevant for understanding stereo abstractly, and concretely was highly informative toward the physiology. Although a "stereo columnar architecture" is not obvious from the anatomy, it is well-formed computationally.

Technically, we proposed a sub-Riemannian model on the space of position and orientation $\mathbb{R}^3 \rtimes \mathbb{S}^2$ for the description of the perceptual space of the neural cells involved. This geometrical structure favors the tangent direction of a 3*D* curve stimulus. The integral curves of the sub-Riemannian structure encode the notions of curvature and torsion within their coefficients, and are introduced to describe the connections between elements. This model can be seen as an extension in the three-dimensional scene of the 2-dimensional association field. In particular, the integral curves of the sub-Riemannian structure of the 3*D* space of position-orientation are exactly those that locally correspond to psychophysical association fields.

Although the goal of this paper is not to solve the stereo correspondence problem, we have seen how the geometry we propose is a good starting point to understand how to match left and right points and features. We used binocular receptive fields to prioritize orientation preferences and orientation differences under the assumption that neuronal circuitry has developed to facilitate the interpolation of contours in 3D space. On the other side, the neurogeometrical method has been coupled with a probabilistic methods for example in Sanguinetti et al. (2010) and Sarti and Citti (2015). Here, the authors studied an analogous problem for generation of perceptual units in monocular vision: they introduced stochastic differential equations, analogous to the integral curves of vector fields, and used its probability density as a kernel able to generate monocular perceptual units. In Montobbio et al. (2019), the probability kernel is built in a direction starting from the receptive fields. A future development of the model will consist in adapting the technique of Sarti and Citti (2015) to find the probability of the co-occurrence between two elements, and individuate percepts in 3*D* space. Individuation of percepts through harmonic analysis on the sub-Riemannian structure has been proposed in the past, both for 2*D* spatial stimuli (Sarti and Citti, 2015) and in 2*D* + time spatio-temporal stimuli (Barbieri et al., 2014a). It would be interesting to develop a similar analysis and extend it to stereo vision.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. MB, GC, and AS were supported by EU Project, GHAIA, Geometric and Harmonic Analysis with Interdisciplinary Applications, H2020-MSCA-RISE-2017. SZ was supported in part by US NIH EY031059 and by US NSF CRCNS 1822598.

Acknowledgments

The current article is part of the first named author's Ph.D. thesis (Bolelli, 2023) and a preprint version is available at Bolelli et al. (2023b).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2023.1142621/full#supplementary-material

References

Abbasi-Sureshjani, S., Favali, M., Citti, G., Sarti, A., and ter Haar Romeny, B. M. (2017). Curvature integration in a 5D kernel for extracting vessel connections in retinal images. *IEEE Trans. Image Process.* 27, 606–621. doi: 10.1109/TIP.2017.2761543

Agrachev, A., Barilari, D., and Boscain, U. (2019). *A Comprehensive Introduction to Sub-Riemannian Geometry*. Vol. 181. Cambridge: Cambridge University Press. doi: 10.1017/9781108677325

Alibhai, S., and Zucker, S. W. (2000). "Contour-based correspondence for stereo," in *Computer Vision - ECCV 2000* (Berlin; Heidelberg: Springer), 314–330. doi: 10.1007/3-540-45054-8_21

Anderson, B. L., Singh, M., and Fleming, R. W. (2002). The interpolation of object and surface structure. *Cogn. Psychol.* 44, 148–190. doi: 10.1006/cogp.2001.0765

Anzai, A., Ohzawa, I., and Freeman, R. (1999a). Neural mechanisms for encoding binocular disparity: receptive field position versus phase. *J. Neurophysiol.* 82, 874–890. doi: 10.1152/jn.1999.82.2.874

Anzai, A., Ohzawa, I., and Freeman, R. (1999b). Neural mechanisms for processing binocular information I. Simple cells. *J. Neurophysiol.* 82, 891–908. doi: 10.1152/jn.1999.82.2.891

August, J., and Zucker, S. W. (2000). "The curve indicator random field: curve organization via edge correlation," in *The Kluwer International Series in Engineering and Computer Science* (Boston, MA: Springer US), 265–288. doi: 10.1007/978-1-4615-4413-5_15

Barbieri, D., Citti, G., Cocci, G., and Sarti, A. (2014a). A cortical-inspired geometry for contour perception and motion integration. *J. Math. Imaging Vision* 49, 511–529. doi: 10.1007/s10851-013-0482-z

Barbieri, D., Citti, G., Sanguinetti, G., and Sarti, A. (2012). An uncertainty principle underlying the functional architecture of V1. *J. Physiol.* 106, 183–193. doi: 10.1016/j.jphysparis.2012.03.001

Barbieri, D., Citti, G., and Sarti, A. (2014b). How uncertainty bounds the shape index of simple cells. *J. Math. Neurosci.* 4, 5. doi: 10.1186/2190-8567-4-5

Baspinar, E., Sarti, A., and Citti, G. (2020). A sub-riemannian model of the visual cortex with frequency and phase. *J. Math. Neurosci.* 10, 1–31. doi: 10.1186/s13408-020-00089-6

Ben-Shahar, O., and Zucker, S. W. (2004). Geometrical computations explain projection patterns of long-range horizontal connections in visual cortex. *Neural Comput.* 16, 445–476. doi: 10.1162/089976604772744866

Blasdel, G. (1992). Orientation selectivity, preference, and continuity in monkey striate cortex. *J. Neurosci.* 12, 3139–3161. doi: 10.1523/JNEUROSCI.12-08-03139.1992

Bolelli, M. V. (2023). *Neurogeometry of stereo vision* (theses). Sorbonne Université; Universitá degli studi, Bologne, Italy.

Bolelli, M. V., Citti, G., Sarti, A., and Zucker, S. (2023a). "A neurogeometric stereo model for individuation of 3D perceptual units," in *International Conference on Geometric Science of Information* (Springer), 53–62. doi: 10.1007/978-3-031-38271-0_6

Bolelli, M. V., Citti, G., Sarti, A., and Zucker, S. W. (2023b). Good continuation in 3D: the neurogeometry of stereo vision. *arXiv preprint arXiv:2301.04542*.

Bosking, W. H., Zhang, Y., Schofield, B., and Fitzpatrick, D. (1997). Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *J. Neurosci.* 17, 2112–2127. doi: 10.1523/JNEUROSCI.17-06-02112.1997

Bridge, H., and Cumming, B. G. (2001). Responses of macaque V1 neurons to binocular orientation differences. *J. Neurosci.* 21, 7293–7302. doi: 10.1523/JNEUROSCI.21-18-07293.2001

Bridge, H., Cumming, B. G., and Parker, A. J. (2001). Modeling V1 neuronal responses to orientation disparity. *Visual Neurosci.* 18, 879–891. doi: 10.1017/S0952523801186049

Burge, J. (2020). Image-computable ideal observers for tasks with natural stimuli. *Annu. Rev. Vision Sci.* 6, 491–517. doi: 10.1146/annurev-vision-030320-041134

Burge, J., and Geisler, W. S. (2014). Optimal disparity estimation in natural stereo images. *J. Vision* 14, 1. doi: 10.1167/14.2.1

Cagenello, R., and Rogers, B. J. (1993). Anisotropies in the perception of stereoscopic surfaces: the role of orientation disparity. *Vision Res.* 33, 2189–2201. doi: 10.1016/0042-6989(93)90099-I

Chang, J. T., Whitney, D., and Fitzpatrick, D. (2020). Experience-dependent reorganization drives development of a binocularly unified cortical representation of orientation. *Neuron* 107, 338–350. doi: 10.1016/j.neuron.2020.04.022

Citti, G., Grafakos, L., Pérez, C., Sarti, A., and Zhong, X. (2015). Harmonic and geometric analysis. *Springer*. doi: 10.1007/978-3-0348-0408-0

Citti, G., and Sarti, A. (2006). A cortical based model of perceptual completion in the roto-translation space. *J. Math. Imaging Vision* 24, 307–326. doi: 10.1007/s10851-005-3630-2

Citti, G., and Sarti, A. (2014). *Neuromathematics of Vision*. Vol. 32. Berlin; Heidelberg: Springer. doi: 10.1007/978-3-642-34444-2

Cumming, B. G., and DeAngelis, G. C. (2001). The physiology of stereopsis. *Annu. Rev. Neurosci.* 24, 203–238. doi: 10.1146/annurev.neuro.24.1.203

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am. A* 2, 1160. doi: 10.1364/JOSAA.2.001160

Deas, L. M., and Wilcox, L. M. (2014). Gestalt grouping via closure degrades suprathreshold depth percepts. *J. Vision* 14, 14. doi: 10.1167/14.9.14

Deas, L. M., and Wilcox, L. M. (2015). Perceptual grouping via binocular disparity: the impact of stereoscopic good continuation. *J. Vision* 15, 11. doi: 10.1167/15.11.11

Duits, R., Boscain, U., Rossi, F., and Sachkov, Y. (2013). Association fields via cuspless sub-riemannian geodesics in SE(2). *J. Math. Imaging Vision* 49, 384–417. doi: 10.1007/s10851-013-0475-y

Duits, R., and Franken, E. (2011). Left-invariant diffusions on the space of positions and orientations and their application to crossing-preserving smoothing of hardi images. *Int. J. Comput. Vision*. doi: 10.1007/s11263-010-0332-z

Elder, J. H., and Goldberg, R. M. (2002). Ecological statistics of gestalt laws for the perceptual organization of contours. *J. Vision* 2, 5–5. doi: 10.1167/2.4.5

Faugeras, O. (1993). *Three-Dimensional Computer Vision: A Geometric Viewpoint*. Cambridge, MA; London: MIT Press.

Field, D. J., Hayes, A., and Hess, R. F. (1993). Contour integration by the human visual system: Evidence for a local "association field". *Vision Res.* 33, 173–193. doi: 10.1016/0042-6989(93)90156-Q

Fulvio, J. M., Singh, M., and Maloney, L. T. (2008). Precision and consistency of contour interpolation. *Vision Res.* 48, 831–849. doi: 10.1016/j.visres.2007.12.018

Geisler, W. S., and Perry, J. S. (2009). Contour statistics in natural images: grouping across occlusions. *Visual Neurosci.* 26, 109–121. doi: 10.1017/S0952523808080875

Geisler, W. S., Perry, J. S., Super, B. J., and Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Res.* 41, 711–724. doi: 10.1016/S0042-6989(00)00277-7

Hess, R. F., and Field, D. J. (1995). Contour integration across depth. *Vision Res.* 35, 1699–1711. doi: 10.1016/0042-6989(94)00261-J

Hess, R. F., Hayes, A., and Field, D. J. (2003). Contour integration and cortical processing. *J. Physiol.* 97, 105–119. doi: 10.1016/j.jphysparis.2003.09.013

Hess, R. F., Hayes, A., and Kingdom, F. A. A. (1997). Integrating contours within and through depth. *Vision Res.* 37, 691–696. doi: 10.1016/S0042-6989(96)00215-5

Hess, R. F., May, K. A., and Dumoulin, S. O. (2014). "Contour integration: psychophysical, neurophysiological, and computational perspectives," in *The Oxford Handbook of Perceptual Organization*, ed J. Wagemans (Oxford Academic). doi: 10.1093/oxfordhb/9780199686858.013.013

Hoffman, W. C. (1989). The visual cortex is a contact bundle. *Appl. Math. Comput.* 32, 137–167. doi: 10.1016/0096-3003(89)90091-X

Howard, I. P. (2012). *Perceiving in Depth, Volume 1: Basic Mechanisms*. New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780199764143.001.0001

Howard, I. P., and Rogers, B. J. (1995). *Binocular Vision and Stereopsis*. New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780195084764.001.0001

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154. doi: 10.1113/jphysiol.1962.sp006837

Hubel, D. H., and Wiesel, T. N. (1970). Stereoscopic vision in macaque monkey: cells sensitive to binocular depth in area 18 of the macaque monkey cortex. *Nature* 225, 41–42. doi: 10.1038/225041a0

Intoy, J., Cox, M. A., Alicic, E., Victor, J. D., Banks, M. S., and Rucci, M. (2021). Fixational eye movements contribute to stereopsis. *J. Vision* 21, 2112–2112. doi: 10.1167/jov.21.9.2112

Ivry, R., Beck, J., and Rosenfeld, A. (1989). Line segregation. *Spat. Vision* 4, 75–101. doi: 10.1163/156856889X00068

Jaeger, D., and Ranu, J. (2015). *Encyclopedia of Computational Neuroscience*. New York, NY: Springer. doi: 10.1007/978-1-4614-6675-8

Jaini, P., and Burge, J. (2017). Linking normative models of natural tasks to descriptive models of neural response. *J. Vision* 17, 16–16. doi: 10.1167/17.12.16

Jones, D. G., and Malik, J. (1991). *Determining Three-Dimensional Shape from Orientation and Spatial Frequency Disparities I-Using Corresponding Line Elements*. Technical Report UCB/CSD-91-656, EECS Department, University of California, Berkeley, CA.

Jones, J. P., and Palmer, L. A. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.* 58, 1233–1258. doi: 10.1152/jn.1987.58.6.1233

Julesz, B. (1971). *Foundations of Cyclopean Perception*. Chicago, IL: The University of Chicago Press.

Kato, D., Baba, M., Sasaki, K. S., and Ohzawa, I. (2016). Effects of generalized pooling on binocular disparity selectivity of neurons in the early visual cortex. *Philos. Trans. R. Soc. B Biol. Sci.* 371, 20150266. doi: 10.1098/rstb.2015.0266

Kellman, P. J., Garrigan, P., and Shipley, T. F. (2005a). Object interpolation in three dimensions. *Psychol. Rev.* 112, 586–609. doi: 10.1037/0033-295X.112.3.586

Kellman, P. J., Garrigan, P., Shipley, T. F., Yin, C., and Machado, L. (2005b). 3-D interpolation in object perception: evidence from an objective performance paradigm. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 558–583. doi: 10.1037/0096-1523.31.3.558

Khuu, S. K., Honson, V., and Kim, J. (2016). The perception of three-dimensional contours and the effect of luminance polarity and color change on their detection. *J. Vision* 16, 31. doi: 10.1167/16.3.31

Koenderink, J. J., and van Doorn, A. J. (1987). Representation of local geometry in the visual system. *Biol. Cybernet.* 55, 367–375. doi: 10.1007/BF00318371

Koffka, K. (1963). *Principles of Gestalt Psychology*. New York, NY: A Harbinger Book.

Kohler, W. (1967). Gestalt psychology. *Psychol. Forschung* 31, 18–30. doi: 10.1007/BF00422382

Lawlor, M., and Zucker, S. W. (2013). "Third-order edge statistics: contour continuation, curvature, and cortical connections," in *Advances in Neural Information Processing Systems 26*.

LeVay, S., Hubel, D. H., and Wiesel, T. N. (1975). The pattern of ocular dominance columns in macaque visual cortex revealed by a reduced silver stain. *J. Comp. Neurol.* 159, 559–575. doi: 10.1002/cne.901590408

Li, G., and Zucker, S. W. (2003). "A differential geometrical model for contour-based stereo correspondence," in *Proc. of IEEE Workshop on Variational, Geometric and Level set Methods in Computer Vision* (Nice).

Li, G., and Zucker, S. W. (2006). Contextual inference in contour-based stereo correspondence. *Int. J. Comput. Vision* 69, 59–75. doi: 10.1007/s11263-006-6853-9

Li, G., and Zucker, S. W. (2008). Differential geometric inference in surface stereo. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 72–86. doi: 10.1109/TPAMI.2008.270

Malach, R., Amir, Y., Harel, M., and Grinvald, A. (1993). Relationship between intrinsic connections and functional architecture revealed by optical imaging and *in vivo* targeted biocytin injections in primate striate cortex. *Proc. Natl. Acad. Sci. U.S.A.* 90, 10469–10473. doi: 10.1073/pnas.90.22.10469

Marr, D., and Poggio, T. (1979). A computational theory of human stereo vision. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* 204, 301–328. doi: 10.1098/rspb.1979.0029

Menz, M. D., and Freeman, R. D. (2004). Functional connectivity of disparity-tuned neurons in the visual cortex. *J. Neurophysiol.* 91, 1794–1807. doi: 10.1152/jn.00574.2003

Miolane, N., and Pennec, X. (2016). "A survey of mathematical structures for extending 2D neurogeometry to 3D image processing," in *Medical Computer Vision: Algorithms for Big Data: International Workshop, MCV 2015* (Munich: Springer), 155–167. doi: 10.1007/978-3-319-42016-5_15

Mitchison, G. J., and McKee, S. P. (1990). Mechanisms underlying the anisotropy of stereoscopic tilt perception. *Vision Res.* 30, 1781–1791. doi: 10.1016/0042-6989(90)90159-I

Montobbio, N., Citti, G., and Sarti, A. (2019). From receptive profiles to a metric model of V1. *J. Comput. Neurosci.* 46, 257–277. doi: 10.1007/s10827-019-00716-6

Neilson, P., Neilson, M., and Bye, R. (2018). A riemannian geometry theory of three-dimensional binocular visual perception. *Vision* 2, 43. doi: 10.3390/vision2040043

Nelson, J., Kato, H., and Bishop, P. O. (1977). Discrimination of orientation and position disparities by binocularly activated neurons in cat straite cortex. *J. Neurophysiol.* 40, 260–283. doi: 10.1152/jn.1977.40.2.260

Oluk, C., Bonnen, K., Burge, J., Cormack, L. K., and Geisler, W. S. (2022). Stereo slant discrimination of planar 3d surfaces: Frontoparallel versus planar matching. *J. Vision* 22, 6–6. doi: 10.1167/jov.22.5.6

Parent, P., and Zucker, S. W. (1989). Trace inference, curvature consistency, and curve detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 11, 823–839. doi: 10.1109/34.31445

Parker, A. J., Smith, J. E. T., and Krug, K. (2016). Neural architectures for stereo vision. *Philos. Trans. R. Soc. B Biol. Sci.* 371, 20150261. doi: 10.1098/rstb.2015.0261

Petitot, J. (2008). *Neurogéométrie de la vision: modeles mathematiques et physiques des architectures fonctionnelles*. Palaiseau: Editions Ecole Polytechnique.

Petitot, J., and Tondut, Y. (1999). Vers une neurogéométrie. Fibrations corticales, structures de contact et contours subjectifs modaux. *Mathématiques et Sciences humaines* 145, 5–101. doi: 10.4000/msh.2809

Poggio, G. F. (1995). Mechanisms of stereopsis in monkey visual cortex. *Cereb. Cortex* 5, 193–204. doi: 10.1093/cercor/5.3.193

Read, J. C. (2015). "Stereo vision, models of," in *Encyclopedia of Computational Neuroscience*, eds D. Jaeger and R. Jung (New York, NY: Springer), 2873–2881. doi: 10.1007/978-1-4614-6675-8_560

Read, J. C., and Cumming, B. G. (2007). Sensors for impossible stimuli may solve the stereo correspondence problem. *Nat. Neurosci.* 10, 1322–1328. doi: 10.1038/nn1951

Samonds, J. M., Potetz, B. R., Tyler, C. W., and Lee, T. S. (2013). Recurrent connectivity can account for the dynamics of disparity processing in V1. *J. Neurosci.* 33, 2934–2946. doi: 10.1523/JNEUROSCI.2952-12.2013

Sanguinetti, G., Citti, G., and Sarti, A. (2010). A model of natural image edge co-occurrence in the rototranslation group. *J. Vision* 10, 37. doi: 10.1167/10.14.37

Sarti, A., and Citti, G. (2015). The constitution of visual perceptual units in the functional architecture of V1. *J. Comput. Neurosci.* 38, 285–300. doi: 10.1007/s10827-014-0540-6

Sarti, A., Citti, G., and Petitot, J. (2007). The symplectic structure of the primary visual cortex. *Biol. Cybernet.* 98, 33–48. doi: 10.1007/s00422-007-0194-9

Sarti, A., Citti, G., and Piotrowski, D. (2019). Differential heterogenesis and the emergence of semiotic function. *Semiotica* 2019, 1–34. doi: 10.1515/sem-2018-0109

Sasaki, K. S., Tabuchi, Y., and Ohzawa, I. (2010). Complex cells in the cat striate cortex have multiple disparity detectors in the three-dimensional binocular receptive fields. *J. Neurosci.* 30, 13826–13837. doi: 10.1523/JNEUROSCI.1135-10.2010

Schmidt, K. E., Goebel, R., Löwel, S., and Singer, W. (1997). The perceptual grouping criterion of colinearity is reflected by anisotropies of connections in the primary visual cortex. *Eur. J. Neurosci.* 9, 1083–1089. doi: 10.1111/j.1460-9568.1997.tb01459.x

Scholl, B., Tepohl, C., Ryan, M. A., Thomas, C. I., Kamasawa, N., and Fitzpatrick, D. (2022). A binocular synaptic network supports interocular response alignment in visual cortical neurons. *Neuron* 110, 1573–1584. doi: 10.1016/j.neuron.2022.01.023

Schreiber, K. M., Hillis, J. M., Filippini, H. R., Schor, C. M., and Banks, M. S. (2008). The surface of the empirical horopter. *J. Vision* 8, 7–7. doi: 10.1167/8.3.7

Singh, M., and Fulvio, J. M. (2005). Visual extrapolation of contour geometry. *Proc. Natl. Acad. Sci. U.S.A.* 102, 939–944. doi: 10.1073/pnas.0408444102

Singh, M., and Fulvio, J. M. (2007). Bayesian contour extrapolation: geometric determinants of good continuation. *Vision Res.* 47, 783–798. doi: 10.1016/j.visres.2006.11.022

Smits, J. T. S., and Vos, P. G. (1987). The perception of continuous curves in dot stimuli. *Perception* 16, 121–131. doi: 10.1068/p160121

Ts'o, D. Y., Zarella, M., and Burkitt, G. (2009). Whither the hypercolumn? *J. Physiol.* 587, 2791–2805. doi: 10.1113/jphysiol.2009.171082

Tu, L. W. (2011). *An Introduction to Manifolds*. New York, NY: Springer. doi: 10.1007/978-1-4419-7400-6

Uttal, W. R. (1983). *Visual Form Detection in 3-Dimensional Space*. Hillsdale, NJ: L. Erlbaum Associates.

Uttal, W. R. (2013). *Visual Form Detection in Three-Dimensional Space*. New Jersey, NJ: Psychology Press. doi: 10.4324/9780203781166

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., et al. (2012). A century of gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychol. Bull.* 138, 1172. doi: 10.1037/a0029333

Warren, P. A., Maloney, L. T., and Landy, M. S. (2002). Interpolating sampled contours in 3-D: analyses of variability and bias. *Vision Res.* 42, 2431–2446. doi: 10.1016/S0042-6989(02)00266-3

Zucker, S. (2006). "Differential geometry from the frenet point of view: boundary detection, stereo, texture and color," in *Handbook of Mathematical Models in Computer Vision* (Boston, MA: Springer), 357–373. doi: 10.1007/0-387-28831-7_22

Zucker, S. W. (2014). Stereo, shading, and surfaces: curvature constraints couple neural computations. *Proc. IEEE* 102, 812–829. doi: 10.1109/JPROC.2014.2314723

Zucker, S. W., and Davis, S. (1988). Points and endpoints: a size/spacing constraint for dot grouping. *Perception* 17, 229–247. doi: 10.1068/p170229

frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 16 May 2024
DOI 10.3389/fcomp.2024.1367534

OPEN ACCESS

EDITED BY
Dirk Bernhardt-Walther,
University of Toronto, Canada

REVIEWED BY
Walter Gerbino,
University of Trieste, Italy
Thiago Leiros Costa,
KU Leuven, Belgium

*CORRESPONDENCE
Nicholas Baker
✉ nbaker1@luc.edu

RECEIVED 08 January 2024
ACCEPTED 03 April 2024
PUBLISHED 16 May 2024

CITATION
Baker N and Kellman PJ (2024) Shape from dots: a window into abstraction processes in visual perception.
Front. Comput. Sci. 6:1367534.
doi: 10.3389/fcomp.2024.1367534

Shape from dots: a window into abstraction processes in visual perception

Nicholas Baker[1]* and Philip J. Kellman[2]

[1]Department of Psychology, Loyola University Chicago, Chicago, IL, United States, [2]Department of Psychology, University of California, Los Angeles, Los Angeles, CA, United States

Introduction: A remarkable phenomenon in perception is that the visual system spontaneously organizes sets of discrete elements into abstract shape representations. We studied perceptual performance with dot displays to discover what spatial relationships support shape perception.

Methods: In Experiment 1, we tested conditions that lead dot arrays to be perceived as smooth contours vs. having vertices. We found that the perception of a smooth contour vs. a vertex was influenced by spatial relations between dots beyond the three points that define the angle of the point in question. However, there appeared to be a hard boundary around 90° such that any angle 90° or less was perceived as a vertex regardless of the spatial relations of ancillary dots. We hypothesized that dot arrays whose triplets were perceived as smooth curves would be more readily perceived as a unitary object because they can be encoded more economically. In Experiment 2, we generated dot arrays with and without such "vertex triplets" and compared participants' phenomenological reports of a unified shape with smooth curves vs. shapes with angular corners. Observers gave higher shape ratings for dot arrays from curvilinear shapes. In Experiment 3, we tested shape encoding using a mental rotation task. Participants judged whether two dot arrays were the same or different at five angular differences. Subjects responded reliably faster for displays without vertex triplets, suggesting economical encoding of smooth displays. We followed this up in Experiment 4 using a visual search task. Shapes with and without vertex triplets were embedded in arrays with 25 distractor dots. Participants were asked to detect which display in a 2IFC paradigm contained a shape against a distractor with random dots. Performance was better when the dots were sampled from a smooth shape than when they were sampled from a shape with vertex triplets.

Results and discussion: These results suggest that the visual system processes dot arrangements as coherent shapes automatically using precise smoothness constraints. This ability may be a consequence of processes that extract curvature in defining object shape and is consistent with recent theory and evidence suggesting that 2D contour representations are composed of constant curvature primitives.

KEYWORDS

perceptual organization, shape, Gestalt, contour perception, abstraction, curvature, visual perception

Introduction

Among the most useful functions of the visual system is the perception and representation of shape. A striking and revealing example is the spontaneous perception of a unified shape from disconnected dot elements. Consider, for example, the array of dots presented in Figure 1A. Although the dots are disconnected and the shape unfamiliar, a well-defined,

coherent shape is spontaneously perceived. Organization into a configural whole does not depend on similarity in the elements' size, color, or element shapes, as shown in Figures 1B–D. Unlike displays used in research on path integration (e.g., Field et al., 1993; c.f., Kellman and Fuchser, 2023), the dot elements have no explicit orientation. The visual system could, in principle, interpolate any number of possible contours between dots in this array (c.f., Kanizsa, 1979).

A familiar example of observers spontaneously perceiving shape from separated points in space is the perception of constellations in the night sky (Metzger, 2009). These organizations turn out to be surprisingly consistent across cultures (Kemp et al., 2022), suggesting that the extraction of shapes from stars depends on basic processes of human perception (Kelly et al., 2024). There also seems to be a high degree of consistency in the shapes observers extract from disconnected dots sampled from novel contours.

Although these observations are commonplace, they are remarkable manifestations of processes of abstraction in visual perception. Baker and Kellman (2018) found that brief exposures of dot arrays to human observers produced perceptual representations that readily supported matching of shape across transformations of position, scale, and orientation. They also found that, even when tasked exclusively with trying to detect changes in the positions of dots, observers had no ability to distinguish changes in dot positions (across two exposures) when dots were moved along a never-shown virtual contour from which the first array of dots was sampled (see Figure 2). These findings provided evidence that perception of these displays produced abstract shape representations, not tied to the particular stimulus elements presented. Such representations are extracted from relations of stimulus elements, but those specific elements are only transiently encoded. Metzger (2009) likewise observed that viewers exposed to a pattern of dots will often substitute the shape defined by the dots in visual memory, and these observations are consistent with many other demonstrations about perception of shape by Gestalt psychologists (see Koffka, 1935 for a review).

Shape perception from dots comprises an especially valuable example of the abstract, relational character of visual perception (Baker and Kellman, 2018; c.f., Kellman and Massey, 2013). Because the notion of abstraction has been used in diverse ways in cognition and perception (e.g., Barsalou, 2003), it is reasonable to ask what we mean in describing a perceptual representation as abstract or a perceptual process as performing abstraction. Baker et al. (2021) suggested that three criteria characterize abstraction in perceptual encoding. Perceptual representations are abstract when they are: (1) *relational*, such that the relevant information encoded, as in the case of shape, is defined over, but not by, constituent elements; (2) *economical*, in that they involve summary descriptions from which much information relating to specific stimulus elements has been discarded; and (3) *additive*, in that abstract perceptual representations may add information that was not strictly in the stimulus information given. Perception of shape from dots exhibits all of these properties. From the relations of dots, a shape representation is obtained that transfers across changes in elements, scale, orientation, etc.; the stimulus elements themselves are not durably encoded, and in fact, many different sets of elements may give rise to the same abstract representation; and the shape representation itself has continuity of contour and a unity that is not given in the stimulus. Recognizing that abstraction—as indicated by these properties—is pervasive in perception may be valuable in clarifying a number of issues in both classical views of perception and recent proposals (e.g., Barsalou, 1999) about the relation between perception and cognition (Kellman and Massey, 2013).

Arguably, perceiving shape from contours that are not made from dots, i.e., that are continuous in the projection to the eyes, also involves all of these same characteristics. If an early level of cortical processing encodes the input into activations of cortical units having orientation sensitivity in local receptive fields, we encounter the same issues of how numerous local activations become organized into tokens of continuous contours and well-defined shape. When the stimulus itself has continuity, it may be harder to realize that an abstract re-description of the input occurs in those cases as well. Shape from dots provides a unique window into these processes, both intuitively and experimentally, as it is easier to point to aspects of perceptual representations that do not exist in the stimulus.

As remarkable as it is that the visual system encodes shape representations from unconnected dots, not all dot arrays give rise to a shape percept. Why do some spatial relations among dots result in a configural whol while others do not? Experiments testing dot perception with a highly constrained number of dots have found that certain relations between dots result in *emergent features*, relations among groups of dots that are more salient than the sum of the dot's individual properties. Pomerantz and Portillo (2011) studied how different spatial relations among two to four dots influence the perception of emergent features in a dot array. They found that observers are highly sensitive to orientation and proximity relations between pairs of dots (see also Hawkins et al., 2016). These relations were shown to influence performance in an odd-one-out task even as differences between the target and distractor approached the minimum threshold for detection (Costa and Wagemans, 2021). Higher-order relations among dots resulted

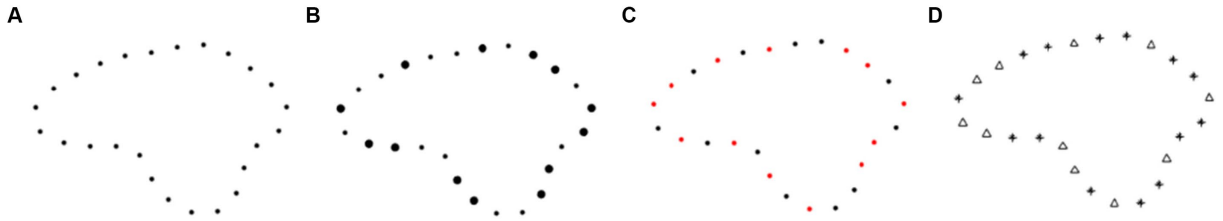


FIGURE 1
Different dot displays giving rise to the same perceived shape. Perceivers organize the elements in each of the four displays into a similar unified shape representation despite variations in element size, color, and shape. **(A)** A shape defined by uniform elements. **(B)** The same shape defined by elements with nonuniform size. **(C)** The same shape defined by elements with nonuniform color. **(D)** The same shape defined by elements with nonuniform shape.

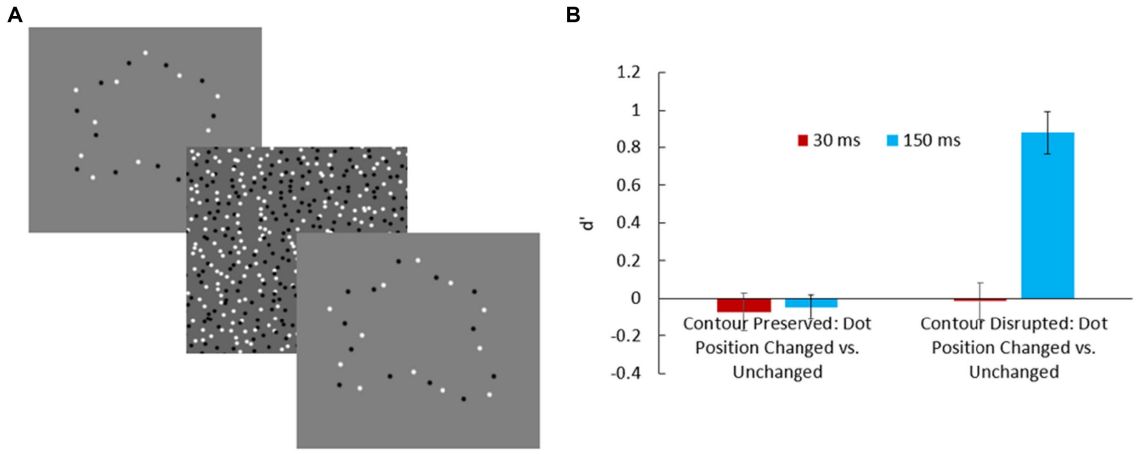


FIGURE 2
(A) Trial from Baker and Kellman (2018) in which dots were shifted along a shape's virtual contour. **(B)** Sensitivity to dot position change. The observer's task was to indicate whether any dots moved between the first and second exposures. When dot positions changed, they could either preserve or disrupt the virtual shape. For 30 ms exposures, there was no sensitivity to dot position changes, indicating that observers did not reliably encode specific dot positions, and, as earlier experiments indicated, 30 ms exposures are too short to allow formation of abstract shape representations. At 150 ms, when dots were shifted along the virtual contour, participants showed no sensitivity to the change; however, dot position changes that disrupted the virtual shape were more detectable, presumably due to their effect on the overall shape representation extracted.

in emergent features less consistently. Dots that are collinear tended to be perceived configurally, but other relations of three dots, such as symmetry, resulted in an emergent feature only for certain displays.

These results provide important insight into how even the simplest displays are perceived configurally, but they cannot fully explain how *shape* emerges from arrays of dots. As Pomerantz and Portillo (2011) point out, the systematic tracking of emergent features becomes difficult or impossible as the number of elements in a display increases. Studies on the percept of shape from randomly placed dots have found that certain configurations are much more frequently perceived to have a unitary shape description than others, depending on proximity and good continuation between the dot elements (van den Berg, 2006). Considerable research has been done on how the visual system organizes an array of dots into multiple distinct shapes, or into a single shape among random dot distractors. Proximity appears to play a major role in these computations (Van Oeffelen and Vos, 1983; Kubovy and Wagemans, 1995; Papari and Petkov, 2005), although similarity (Zucker et al., 1983) and good continuation are also important cues (Smits et al., 1985; Lezama et al., 2016).

Notions of proximity and similarity are perhaps intuitive with respect to dot configurations, but good continuation requires some elaboration. Good continuation in Gestalt psychology has always been somewhat vague: Wertheimer (1923) said that "one knows" what it is, and Kanizsa said that it resists simple definition (Kanizsa, 1979). Other work has provided more rigorous, but highly varied, definitions of good continuation.

A common view is that good continuation depends on the degree to which dots are collinear with each other (Smits et al., 1985; Wouterlood and Boselie, 1992; van den Berg, 2006), which has been shown to aid contour detection using simple segments (Uttal, 1973; Prinzmetal and Banks, 1977). Another definition that has been proposed quantifies good continuation by measuring the degree of symmetry between the middle and first dot and the middle and third dot in a triplet. Sequences of dots with more symmetrical triplets are considered less accidental and are therefore more likely to be organized together (Lezama et al., 2016). Still another view is that the visual system evaluates good continuation in quartets. According such theories, the goodness of dots' continuation is quantified by the degree to which the turn angle between consecutive dots in a quartet remains constant (Feldman, 1997; Kelly et al., 2024). This definition is similar to one proposed by Pizlo et al. (1997) which theorizes that a dot sequence has good continuation if it is *smooth*, which they define as consisting of successive pairs of dots whose orientations are similar to each other. For example, a four-dot sequence, ABCD, is considered smooth if the difference in orientation between AB and BC is small and the difference in orientation between BC and CD is small.

A theme shared among many of these theories is that good continuation is primarily a summation of many local computations (Feldman, 1997; Pizlo et al., 1997; Lezama et al., 2016; Kelly et al., 2024). This is consistent with path integration work done using oriented Gabors (Field et al., 1993; Hess and Field, 1999). However, others have argued that good continuation is a more global consideration and is strengthened by monotonicity and curvature regularity along the extent of the dot sequence (Smits and Vos, 1986; Yuen et al., 1990; van den Berg, 2006).

Our goal is not to test these competing formulations of good continuation but to better specify how the manipulation of spatial relations between dots influences. Understanding this could provide crucial insight into how the visual system forms abstract shape representations. This effort is specifically relevant to understanding how contours are formed. Dot arrays, with no continuous contour physically given, may provide unique insight into underlying visual processes in shape representation (Baker and Kellman, 2018).

One way the visual system could form a contour representation from unconnected dots is by interpolating straight edges between adjacent boundary points (e.g., O'Callaghan, 1974). Under this view, dot displays with longer sequences of collinear dots would be simplest because a new line segment would be required for any change in

orientation between sequential dot pairs. It would also be consistent, within a certain size range, with modest activation of large, orientation-sensitive units in visual cortex by pairs or sets of dots. The straight line interpolation process is computationally simple and would effectively minimize the total length of the contour connecting the dots.

On the other hand, we have the phenomenological experience of perceiving smooth curves in displays like the dots in Figure 1. Moreover, connecting straight edges at each point would result in a very jagged percept of the dot array, albeit one that is invariant to rigid 2D transformations. Under this view, any dot between two other dots, A and B, would either need to be collinear with A and B or would be perceived as a corner in the contour representation of the display. Attneave (1954), in classic work, proposed that corners are the positions on the contour with the greatest informational content and are consequently more salient in our perception (Troncoso et al., 2005; De Winter and Wagemans, 2008). Previously, we shifted all dots in an array along a virtual contour, which would result in a contour representation with a completely different set of corners if observers were indeed encoding the shape with straight line interpolation. Given the perceptual importance of corners, such a change should be easily detected, but we found that humans had no sensitivity at all to the positional shift. This suggests that the visual system does not form shape representations of dot arrays by simply interpolating straight lines between adjacent points. Curved edges may instead be interpolated across spans of dots such that shifts to any other position on the interpolated curve are largely undetectable.

There is other empirical evidence that the visual system does not always represent dots as corners in a contour. Koffka (1931) placed points along a circle to estimate the number of dots at which the virtual contour was perceived as smooth. He estimated that this transition occurred at around eight evenly spaced dots (which produces inclusive angles of 135° for each dot triplet). Bouma (1976) gave a more conservative estimate that 10 dots were needed for the virtual contour to be perceived as smooth. Smits and Vos (1987) used more systematic tests to estimate this transition point and found that when the inclusive angle between triplets of dots was sufficiently large (greater than about 140°), the dots were perceived as curvilinear (Figure 3A). In a later study, van Assen and Vos (1999) developed a more objective measure of perceived curvilinearity by measuring participants' bias to say whether a target dot was below or above a virtual contour defined by four other dots. They found that when the inclusive angle between the central dots was 135° to 150°, participants' bias was consistent with perceiving a curved contour. Feldman (1996) systematically varied the inclusive angle for three dot displays and found that the 50% threshold for curvilinear responses vs. angular responses was at around 120°. He found that when a fourth dot was added to the configuration to create two similar angles from dot triplets, the threshold went down (Feldman, 1997; Figure 3B).

If forming a shape representation by encoding each dot position as a corner is both parsimonious and computationally simple, why would the visual system ever extract contour representations with smooth curves from dot arrays? One possibility is that the visual system is sensitive to curvature-specific contour segments and can encode curvilinear contours from a dot array as easily as straight lines (Smits and Vos, 1986; Yuen et al., 1990). Past research on connected contours has shown that perceptual tasks requiring a shape representation are accomplished better and more quickly with smooth contours than with angular contours (Bertamini et al., 2019). Other work has shown that visual system has special facility for encoding contours of constant curvature (Baker et al., 2021). A perceptual corner has a first-order (tangent) discontinuity at its vertex, meaning there will always be a segment boundary at that point. On the other hand, if the contour is perceptually smooth at that point, the entire segment could be represented as a single curvature segment. As a consequence, representing dot patterns' shape with curvilinear segments may be computationally simpler because the shape description consists of fewer parts. Moreover, the presence of a corner or first-order discontinuity, in this case an L-junction, not only forms a boundary within a single contour or object representation, but under some circumstances in visual perception plays a key role in determining that intersecting contours belong to different objects (Clowes, 1971; Shipley and Kellman, 1990; Heitger et al., 1998; Kalar et al., 2010).

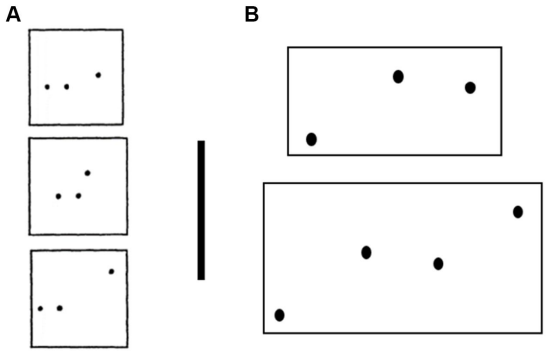


FIGURE 3
(A) Dot triplets used by Smits and Vos (1987). **(B)** Comparison of a dot triplet and quadruplet from Feldman (1997). Feldman found that the addition of a fourth dot with similar turn angle increased the probability of a curvilinear response. Reprinted with permission from *Perception* and *Vision Research*, respectively.

These considerations suggest that the abstract shape representation that is ultimately encoded by an array of dots might be stored more efficiently as a set of relatively few curved segments than a larger set of straight segments. If this is the case, we expected that it would be easier to encode arrangements of dots that appear to have few or no sharp corners as a shape than arrangements of dots with many perceived corners. A different framing of the question is that both smooth contours and first-order contour discontinuities exist in the world and are important to encode. Discrete dots provide sparse information about what structure might best represent their relations. Under what conditions does the visual system encode smooth continuation through an element vs. representing that element as a contour junction?

We studied information that might be used by the visual system to represent smooth curves vs. corners and tested its effects on shape encoding, using both subjective and objective measures. In Experiment 1, we experimentally estimated the angle at which dots in an array are perceived to be on corners vs. smooth curves in a virtual contour, both in conditions with only three points and in conditions where other configural aspects of the display facilitate or inhibit the perception of a corner. In Experiment 2, we used these estimates to create larger dot arrays and asked participants to judge how much

each dot array looked like a coherent shape. In Experiment 3, we showed pairs of dot arrays sampled from smooth or corner shapes at different orientations and measured the time it took for participants to judge whether the shape formed by the dots was the same or different. In Experiment 4, we hid dot arrays sampled from both kinds of shapes among randomly placed distractor dots and measured participants' ability to detect the shape in both conditions.

Experiment 1

The general goals of our study were to understand the conditions under which arrangements of dots are perceived as smooth curves vs. corners and to test the consequences of curve vs. corner encoding on shape perception and representation. As a starting point, we first needed to determine the angles at which a dot is perceived to be on a corner (i.e., a first order discontinuity) vs. when the dot is perceived to be on a smoothly changing edge (i.e., a differentiable point on the contour). Previous work suggests that the threshold between corners and smooth edges is between 120° and 150° (Koffka, 1931; Bouma, 1976; Smits and Vos, 1987; Van Assen and Vos, 1999).

In Experiment 1, we sought to replicate these findings. Because the displays we intended to use in subsequent experiments involved many more dots than the three that define a single vertex, we also wished to test whether the perception of curves vs. corners depended only on the local geometry among dot triplets or if dots more remote from the potential vertex also influenced perception of corners.

Methods

Participants

Thirty (23 female, seven male, $M_{age} = 20.5$) participants from the University of California, Los Angeles participated in this study for course credit. All participants had normal or corrected-to-normal vision.

Display and apparatus

Subjects were seated 70 cm from a 20-inch View Sonic Graphic Series G225f monitor. The monitor was set to a resolution of 1024 × 768 pixels with a refresh rate of 100 hz. Except when noted otherwise, all aspects of the displays and apparatus in subsequent experiments were the same as in Experiment 1.

Stimuli

We had three stimulus categories: convex, concave-convex-concave, and convex-convex-convex, referring to the direction of curvature of the second dot in a sequence of triplets. Convex vs. concave was defined by reference to the upward direction of visual field, so that dot triplets were considered convex if their central dot was above its flankers and concave if it was below its flankers. The "convex" condition had the simplest displays, with a central dot flanked by two lower dots on either side. (This base of three dots was always convex upward.) The distance between the central dot and its two flankers was kept constant, but the position of the flankers could change to manipulate the angle defined by the dot triplet with the central dot as a vertex. These angles were not predetermined but increased or decreased depending on participants' responses.

In the "concave-convex-concave" condition, the central triplet was flanked by one additional dot on each side. The contour defined by the central dot and its two flankers on each side bent upward such that the central dot appeared to be the joining point between two concave edges (see Figure 4 for examples). The two concave edges could rotate around the central dot to increase or decrease the angle between the central dot and its two closest flankers. The convex-convex-convex condition was like the concave-convex-concave condition except that the contour defined by the central dot and its two flankers on either side appeared to be bending downward so that the two edges meeting at the join point were both convex (as viewed from above). This created an array of dots where the direction of curvature (as defined by the turning angle between dot) for the central point matched the direction of curvature for the two points on either side of it. The opposite was true for the concave-convex-concave condition, where the direction of curvature alternated twice. For both five-dot conditions, the arms on either side of the central dot were symmetrical and the angle centered on the second and fourth dot (from left to right—see Figure 4) was fixed at 135°. In the first presentation of the stimulus in any of these conditions, the angle between the central dot (the red dot in Figure 4) and the dots directly to its left and right was 90° (Figure 4, botttom). Subsequent angles were determined by participants' responses.

Dots were evenly spaced 1.8° of visual angle apart from each other. Each dot in the arrangement subtended 7.2 arcmin and the maximum total height or width of a dot arrangement was 9.6° of visual angle. Dots were rendered black ($luma = 0$) of the on a gray background ($luma = 100$).

Design

The experiment consisted of three conditions corresponding to the three stimulus categories (convex, concave-convex-concave, and convex-convex-convex). Trials were interleaved among the three conditions, with the aim of determining the threshold for seeing a corner in each. We used interleaved staircase procedures to determine participants' 50% probability of responding "corner" for each of the three conditions, manipulating the angle between the central dot and its two flankers. Participants completed at least 24 trials per condition. If their responses converged to a 50% threshold after 24 trials, they ceased to see trials for that condition. If their responses had not converged, they continued seeing trials until we found at least three crossover points (Leek, 2001).

Procedure

In each trial, participants were shown a fixation cross followed by a single arrangement of dots in the center of their screen. The dot arrangement could be from any of the three stimulus categories. Participants were instructed that they would be shown a group of dots on the screen in each trial, with a central dot and 1 or 2 dots on either side of it. Participants' instructions were to look at the display and imagine that the dots are connected in some way. They were then told that "Your task is to decide whether the middle dot appears to be a corner or not." They were told to respond "curve" if the middle dot appeared to be on a smooth, curved edge, and "corner" if it appeared to be a pointy feature. Based on pilot work, we found that the correct task could be communicated effectively through this combination of instructing subjects to imagine that the dots are connected in some way and also stressing that they should

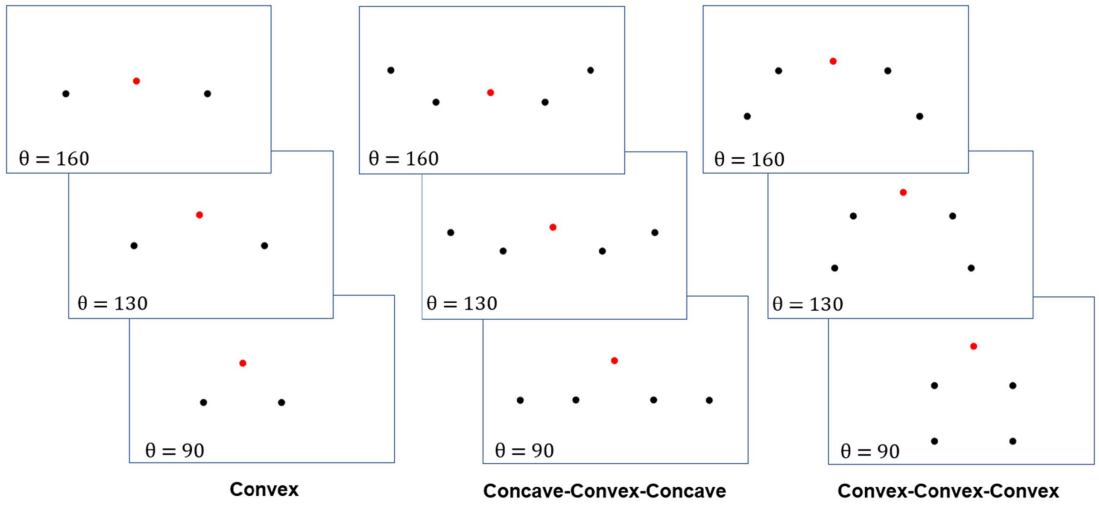


FIGURE 4
Sample displays from Experiment 1. Left: Convex condition. Middle: Concave-Convex-Concave condition. Right: Convex-Convex-Convex condition. The values of θ shown in the bottom left of each display correspond to the angle between the central dot and its nearest flanker on either side. The central dot is highlighted in red only for presentation purposes: it was black in the displays shown to be participants.

indicate how the middle dot *appeared*. Together, these instructions overcame confusion that some participants had in pilot work about instructions to respond based on a perceived virtual contour. Participants were told that there was no expectation that half the trials should be smooth and half should be corners. If they saw the middle dot as the vertex of a corner in all trials, they should say so, and likewise if they saw the middle dot as lying on a smooth curve in all trials.

We used an adaptive staircase procedure, starting with a central angle of 90° for each of the three conditions. After participants responded, we adjusted the angle between the central dot and its flankers based on their response. For the first eight trials in each stimulus category, we adjusted the angle by a larger amount, increasing the angle between the three central dots by 11° if participants reported seeing a corner or decreasing it by 11° if participants reported seeing a smooth curve. In the next eight trials for each stimulus category, the angle increased or decreased by 4.5° depending on participants' responses. After participants had completed 16 trials in a condition, the angle increased or decreased by 1.2° based on participants' responses for all remaining trials. This approach was used to adjust quickly early to get near to participants' 50% threshold, and then to make smaller changes to get a more precise estimation of participants' true threshold. Staircases for each of the three conditions were interleaved to minimize any carryover effects that might occur from sequences of trials with changing angles in a single condition.

A condition ended when participants switched from reporting a curved percept to reporting a corner percept (or vice versa) three times or when they had completed 24 trials of the condition, whichever came later.

Dependent measures and data analysis

We measured the threshold at which participants reported seeing a corner vs. a smooth curve equally often in each of the three conditions. Our expectation was that the threshold for the convex condition (consisting of only three dots) would be between 135° and 150°, consistent with previous findings (Koffka, 1931; Bouma, 1976; Smits and Vos, 1987; Van Assen and Vos, 1999). We predicted that if dots beyond the central three dots influence the perception of a corner vs. a smooth curve, then the concave-convex-concave condition would have a larger threshold than the convex condition because participants would be more likely to perceive a first order discontinuity at the middle point. By the same token, we predicted that the convex-convex-convex condition would have a lower angular threshold because the series of vertices would be consistent with a monotonically curved contour.

Results

The results of Experiment 1 are shown in Figure 5. A one way repeated measures ANOVA confirmed a significant main effect for dot arrangement condition, $F(2,58) = 32.99$, $p < 0.001$, $\eta^2_{partial} = 0.53$. The 50% threshold for three points (the convex condition) closely matches earlier findings. In our experiment, participants' mean threshold was 140°, consistent with the 120° to 150° range reported in previous studies. The concave-convex-concave condition, which we hypothesized would induce more corner percepts, had a mean threshold of 148°. This threshold was not significantly different from the threshold estimated in the convex condition, $t(29) = 0.93$, $p = 0.36$, *Cohen's* $D = 0.17$. The convex-convex-convex condition, which we hypothesized would induce fewer corner percepts, had a mean threshold of 82°. This difference did significantly differ from both the convex condition, $t(29) = 6.18$, $p < 0.001$, *Cohen's* $D = 1.13$ and the concave-convex-concave condition, $t(29) = 7.54$, $p < 0.001$ *Cohen's* $D = 1.38$.

Analysis of the skew and kurtosis of our data suggested that they were not normally distributed, so we applied the Box-Cox transformation (Box and Cox, 1964) and reanalyzed these effects. Analyses of the transformed data found the same general effects as were observed in the

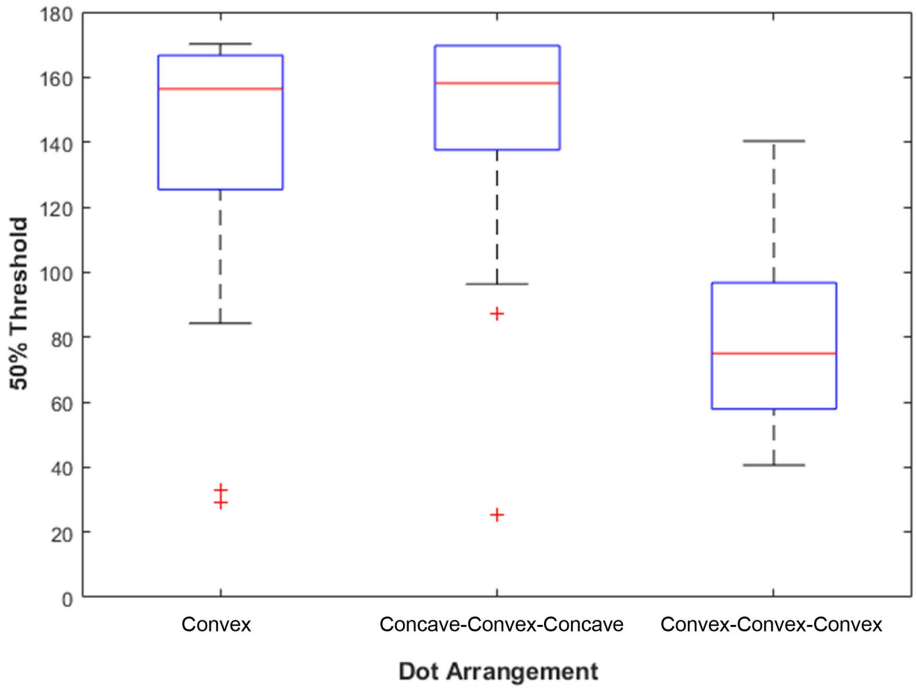


FIGURE 5
Experiment 1 results. The box shows the interquartile range of thresholds for individual participants. The red line shows the sample median for the 50% threshold—the median angle at which participants were as likely to report perceiving the central dot to be on a corner as to report perceiving it on a smooth curve. The "whiskers" extend to the most extreme datapoint within 1.5 times the length of the interquartile range from the top or bottom edge of the box (covering 99.3% of the data if they are normally distributed; McGill et al., 1978; Krzywinski and Altman, 2014). Outliers are data points beyond the whisker and are plotted as red +'s.

untransformed data. A repeated measures ANOVA confirmed significant differences among the three conditions, $F(2,58) = 39.35$, $p < 0.001$, $\eta^2_{partial} = 0.58$ and significant differences were found between the convex and convex-convex-convex conditions [$t(29) = 7.18$, $p < 0.001$, *Cohens' D* = 1.31] but not between the convex and concave-convex-concave conditions [$t(29) = 1.29$, $p = 0.21$, *Cohen's D* = 0.24].

Discussion

The first aim of Experiment 1 was to understand the geometric conditions under which a dot in a sequence of dots would be perceived as a corner. Our data suggest that the threshold for equal probability of perceiving a dot as a corner vs. lying on a smooth contour is around 140°, with angles smaller than that increasingly tending to be seen as corners. The second aim of Experiment 1 was to understand whether the perception of corners vs. smooth curves was a local computation only (i.e., it depended only on the angle between a central dot and the dots on either side of it in a triplet) or if dots outside of the triplet also played a role. Given that the displays we planned to use in subsequent experiments would include arrays of 25 dots, it was important to know whether the corner percept was a purely local computation among the dots defining the angle or if other dots could shift viewers' judgments by changing the way the dot array was perceived.

We found no statistically reliable difference in participants' thresholds for concave-convex-concave displays that were designed to facilitate the perception of a corner. In these displays, there are at least two ways that participants could interpolate virtual contours between the five presented dots. The first way involves two curvature segments that join at the central dot in the display (Figure 6A). This representation of the virtual contour will almost always result in a perceived corner because the tangent of the contour where the first curvature segment ends is very different from the tangent of the contour where the second curvature segment begins, resulting in a first-order discontinuity. The second way a contour could be interpolated is by organizing the central three dots into a curvature segment and the two flanking dots on either side into other segments (Figure 6B). This representation predicts no difference in threshold between the concave-convex-concave condition and the convex condition (Figure 6C) because the same three dots are encoded as their own chunk. Our results suggest either that participants favored the second organization of dots to the first, grouping the central three dots as one unit and the flanking pairs of dots as separate units or that the perception of a corner depends only on the angle between the vertex and the dot on either side of it.

One reason that the second grouping is preferred could have to do with symmetry. The central three dots in the display were always symmetrical over a vertical axis, while the first three and last three dots were symmetrical over diagonal lines in some displays (see Figure 4, bottom middle). Previous research has found that symmetry is much more likely to be an emergent feature when elements are symmetrical about a cardinal axis than when they are symmetrical about a noncardinal axis (Pomerantz and Portillo, 2011), which may have result in better grouping of the central three dots with each other.

By contrast, the convex-convex-convex arrangement of dots did have a significant influence on observers' tendency to perceive a corner. Observers were significantly more likely to see the central dot

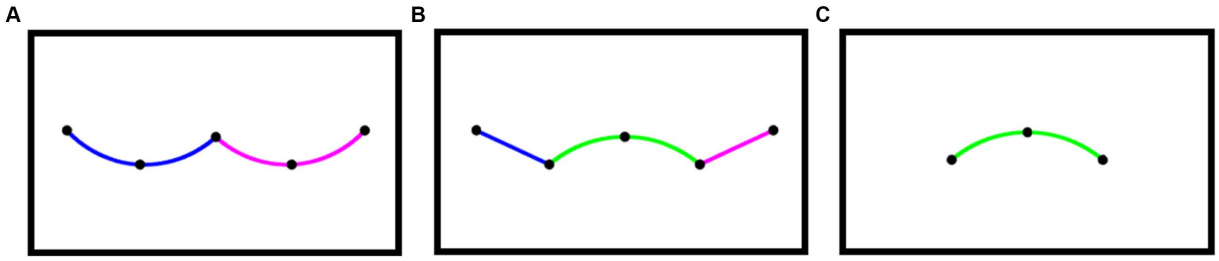


FIGURE 6
Two ways of interpolating contours between dots in the concave condition. **(A)** Two curved segments join at the middle (left). **(B)** Three segments join at the second and fourth dot (middle). **(C)** A curved interpolation between the three central points in **(A,B)** (right). The physical contours chosen to represent the interpolation between dots are chosen arbitrarily in this figure. The edge perceived between dots may be neither a straight edge nor a constant curvature segment. We used these forms of interpolating lines only to help visualize possible organizational structures of the display.

as placed on a smooth curve when additional dots were added that were consistent with a virtual contour with constant curvature polarity. This suggests that there is an asymmetry in the way that additional dots influence our perception of corners. While adding dots to strengthen the percept of a smooth, monotonic curve does weaken the perception of corners, the opposite approach of placing dots so that the contour cannot be monotonic does not appear to have a strengthening effect on the perception of corners.

Gestalt cues shown to produce emergent features like proximity are unlikely to explain differences in the perception of corners in the convex-convex-convex vs. concave-convex-concave conditions. Both displays had identically spaced adjacent dots, but the first and fifth dots in the convex-convex-convex condition were closer to each other than in the concave-convex-concave condition. This difference in proximity may have resulted in a greater overall percept of configural structure in the convex-convex-convex condition, but it is unlikely to have influenced the perception of curvature at the central point. In fact, dots whose extreme points were moved closer together by reducing the angle at the central point were perceived as curvilinear less often than dots whose extreme point remained more distant.

The Gestalt cue likely to be playing the greatest role in observers' perception of a corner vs. a smooth edge is good continuation. As previously discussed, however, there are many different definitions of good continuation in the perception literature, and those that are readily applicable to dots make different predictions about how the five-element dot arrays should be perceived. The data from the convex-convex-convex condition suggest that the perception of good continuation is not solely determined by the local spatial relations in a sequence of dot triplets (e.g., Lezama et al., 2016), but depends on larger clusters of local dot relations (Feldman, 1997; Kelly et al., 2024), or on the global monotonicity of the dot sequence (Smits and Vos, 1986; Yuen et al., 1990).

Experiment 2

In Experiment 1, we studied the tendency of a series of dots to be represented as a connected contour and estimated the threshold, in terms of angular relations, at which a dot in the series is perceived to contain a first order discontinuity. What do these data reflect about perception and representation of contours from sets of separated dots? Experiments 2, 3, and 4 investigated the perceptual reality and impact of perception of seeing a vertex or smooth continuation in sequences of dots. In Experiment 2 we used the thresholds estimated in Experiment 1 to measure the strength of a shape percept in three different kinds of dot displays. We created arrays of dots by sampling from (a) novel shapes with smooth curves; (b) novel shapes with sharp corners, and (c) random dot arrangements. We then asked subjects to rate the degree to which the dots appeared to form a coherent shape. Our prediction was that dots sampled from smooth contours would be judged more shape-like than dots sampled from shapes with corners, which would in turn be judged more shape-like than random dot arrangements.

Methods

Participants

Twenty-five undergraduates (3 male, 22 female, $M_{age} = 20.6$) from the University of California, Los Angeles participated in the study for course credit. All participants had normal or corrected-to-normal vision.

Stimuli

Experiment 2 included three different kinds of dot arrays: Smooth, Corner, and Random. "Smooth" dot arrays were created by placing nine control points at evenly spaced angular positions around a circle and then moving each control point toward or away from the circle's center by a random distance, then fitting cubic splines through the nine control points in polar space (see Figure 7A). The control points' signed displacements were sampled from a normal distribution centered at 0% with a standard deviation of 18%. The mean absolute distance of the control points' displacement was 14.34% (SD = 11%). We made use of our findings from Experiment 1 to create shapes that we expected to generally be perceived as Smooth or Corner shapes. In Experiment 1, we found that when the angle between dots exceeded 140°, the point tended be perceived as smooth, regardless of the spatial relations between dots beyond the triplet determining the angle. In the range between 82° and 140°, the point could be made to appear smoother by adding additional dot triplets whose turn angle had the same polarity as the point in question. The displays used in Experiment 2 were significantly more complex than those used to estimate these thresholds in Experiment 1. They consisted of many more dots and were not symmetrical on either side of any possible vertex. Still, we expected that displays with dot triplets whose angle mostly exceeded 140° would generally be perceived as smooth and

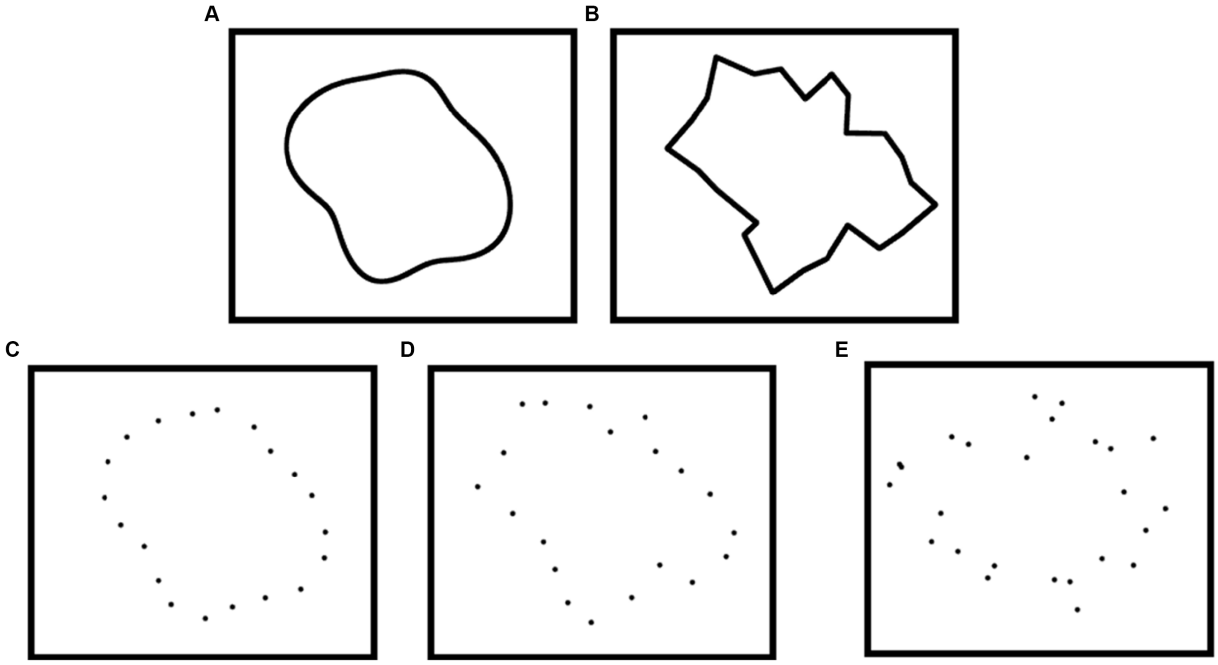


FIGURE 7
Shape stimuli used in Experiment 2. **(A)** A Smooth shape contour. **(B)** A Corner shape contour. **(C)** Dots sampled from the Smooth shape. **(D)** Dots sampled from the Corner shape. **(E)** Random dot arrays.

displays that included dots triplets whose angle was less than 90° would generally be perceived as angular.

We did not directly manipulate the angle between dots in the Smooth condition, but the cubic spline fitting shape generation algorithm produced shapes that were always differentiable at all positions. With nine control points and the possible distances they could be displaced, the angles among dots sampled at these positions tended to be obtuse with an angle that exceeded 140° ($M_{angle} = 154°$, $SD = 16°$). The average minimum angle in Smooth displays was 123° ($SD = 13°$).

From each smooth shape generated as described above, we then sampled 25 points along the contour to get the dot array. The points were sampled nonuniformly by taking 25 evenly sampled points and moving them in a random direction along the contour. Though not directly relevant to this experiment, we included jittering along the contour to prevent participants from using local spatial relationships between a small set of dots rather than the overall shape of a dot array in subsequent experiments using objective performance methods (see Experiment 3 for more explanation). The amount of jitter was randomly sampled from the normal distribution. Dots were shifted along the shape's virtual contour by a random signed distance from a normal distribution, with a mean distance of zero and a standard deviation equal to 4% of the contour's total length. We constrained the display to enforce a minimum distance of 7.2 arcmin between any two points to prevent them from overlapping or appearing to touch each other (Figures 7C,D).

"Corner" dot array stimuli were created by generating a smooth dot array, reducing the angle between some of the dots, fitting straight lines between the set of dots, and then resampling from the straight line contour. To distinguish two closely related concepts here, we refer to a generating figure that consists of all straight line connections between dots as a "cornered figure," and we refer to the resultant derived dot stimuli as "corner stimuli" used in the "Corner condition". We began with a dot array generated by evenly sampling 25 dots from the same kind of shape from which the Smooth dots were sampled. We then altered between 8 and 13 dot triplets in the display (determined for each display by randomly sampling from a uniform distribution of integer values). For a given dot triplet, ABC, we imposed a corner percept by interpolating a line between points A and C, then moving point B perpendicularly away from the interpolated line while simultaneously moving A and C along the line until the vertex at B was between 78° and 90°. This range was chosen so that the Corner shapes would be reliably perceived to have first-order discontinuities based on our Experiment 1 findings while also including some natural variability in the angle between dot triplets. Because the angle reduction process resulted in shapes with much less regular spacing than in the Smooth condition, we interpolated straight lines between the 25 repositioned points to get a new shape contour with corners (Figure 7B). From each such shape, we used the same nonuniform sampling procedure as for smooth dot displays to sample 25 new dots and generate a dot figure (Figure 7D). The resulting dot display had between one and two dot triplets whose angle was less than 90° and several more whose angle was between 90° and 100°.

"Random" dot array stimuli also began with the smooth dot array. Rather than moving dots to reduce the angle between them, dots were moved in random directions. Each of the 25 points was moved a distance equal to the total length of the contour divided by 25 in a random direction, with the constraint that dots could not be closer together than 7.2 arcmin (Figure 7E). We used this method instead of truly random placement to prevent subjects from judging shape based on whether there was an open center within the dot array, a feature

that was naturally present in both the "Smooth" and "Corner" displays. The method we used to generate "Random" dot arrays matched the open center of the other two conditions without appearing to define a shape.

Dot arrays for each of the three conditions were matched in size, subtending on average 12.0° of visual field along the longer of their horizontal and vertical dimension, and at most 20.7° of visual field.

Design

Experiment 2 had three conditions, with 70 trials in each. In the Smooth condition, we showed dot arrays from smooth shapes. In the Corner condition, we showed dot arrays from shapes with corners. In the Random condition, we showed random dot arrays. Trials with all three conditions were randomly interleaved.

Procedure

On each trial, participants were shown one of the three stimulus types and asked to evaluate the degree to which the dot array seemed to form a shape. The dot array remained on the screen until a response was given. Participants were instructed to rate the display on a 6-point scale, ranging from "The dots look totally random" to "The dots look totally like a shape." Integer responses between 1 and 3 reflect that the dots looked more random than like a shape to viewers, either slightly so (3), moderately so (2) or strongly so (1). Integer responses between 4 and 6 reflect that the dots appeared more like a shape than a random set of dots and followed the same progression. We instructed participants to use all 6 response options to reflect qualitative differences in the degree to which different dot arrays appeared to be shapes. Before beginning the main experiment, participants completed five practice trials to familiarize themselves with the response buttons and to view all three stimulus types before giving recorded shape judgments.

Results

Figure 8 shows the primary results of this experiment. Figure 8A shows the mean subjective rating for each of the three stimulus types. There is a clear ordering in which dots sampled from Smooth contours were perceived as most shape-like, followed by dots sampled from cornered contours, followed by randomly sampled dots. This pattern was shown by every participant who completed the experiment. A one-way ANOVA confirmed a significant difference between the groups, $F(2,48) = 681.86$, $p < 0.001$, $\eta^2_{partial} = 0.97$ and Bonferroni-corrected paired sample t-tests confirmed that dots sampled from smooth contours were rated more shape-like than dots sampled from cornered contours, $t(24) = 14.93$, $p < 0.001$, *Cohen's* $d = 2.89$ and that dots sampled from cornered contours were rated more shape-like than randomly sampled dots, $t(24) = 20.64$, *Cohen's* $d = 4.13$ $p < 0.001$.

We also analyzed the average number of trials in which subjects perceived a shape at all. For this measure, we included any display that received a subjective rating greater than 3. The results are shown in Figure 8B. Paired samples t-tests confirmed that subjects' perceived significantly more of the dots sampled from smooth contours as a shape than they did dots sampled from cornered contours, $t(24) = 6.04$, $p < 0.001$, *Cohen's* $d = 1.21$ and that dots sampled from cornered contours were perceived as shapes significantly more often than randomly sampled dots, $t(24) = 13.81$, $p < 0.001$, *Cohen's* $d = 2.76$.

Discussion

Experiment 2 furnished evidence that dots sampled from smooth contours are more phenomenologically shape-like than dots sampled from contours with sharp corners. Every participant gave higher shape ratings for the smooth contour condition and reported perceiving more of the smooth contours as shapes than the cornered contours. Participants never saw the underlying contour from which the dot arrays were sampled. Geometrically, all the dots sampled from shapes with corners could be represented with curvilinear contours, but participants made qualitatively different responses for Corner dot arrays.

One reason that participants may have given lower shape ratings for the Corner stimuli is that they were interpolating a curvilinear contour through the dots in the Corner displays, but the process was more difficult for dot arrays with sharper angles. Though the data cannot rule out this possibility, we consider it unlikely because the

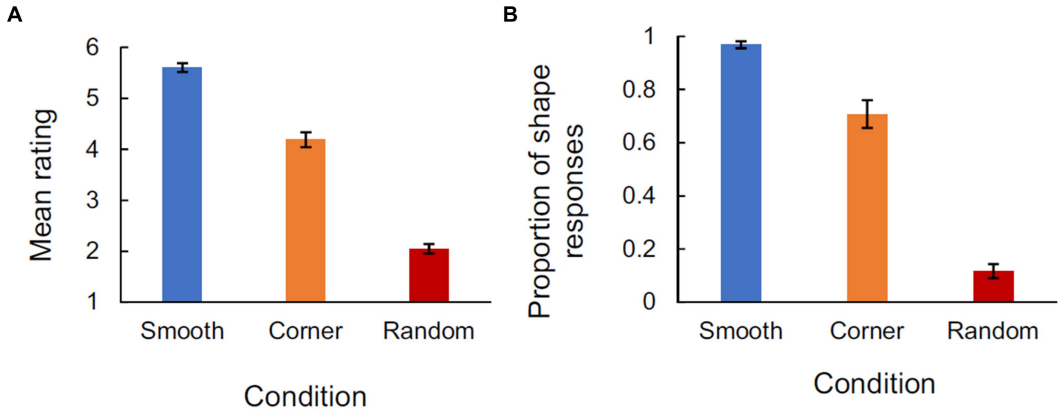


FIGURE 8
Experiment 2 results. **(A)** Participants rating of shape for each of the three conditions. **(B)** The percentage of ratings that were more shape-like (i.e., rating > 3) for each condition.

results of Experiment 1 suggest that we perceive corners at certain points in the Corner displays. Our phenomenological experience of Corner displays like the one in Figure 7D also suggest that the array is perceived with corners.

A more likely possibility is that Corner displays were harder to encode as shapes than Smooth displays because there was more variety in the presented angles. Corner displays were likely perceived as having a mix of smooth curves and corners, whereas the Smooth displays were likely only perceived with smooth curves. As a result, there would be greater perceived homogeneity among dot triplets in the Smooth displays than in the corner displays. The irregularity between smoothly bending and/or straight edges and edges that abruptly change direction could make it more difficult to resolve Corner displays into shapes.

Experiment 1 tested in simple dot arrays the relations that lead to perception of corners. Experiment 2 used the results of Experiment 1 and showed that in more complex arrays, the quantitative estimate of what angular relation produces corner perception in simple arrays also predicts perception of smooth shapes in more complex ones. Both of these experiments, however, used only subjective measures of smoothness or perception of a shape. Subjective methods have a useful role in perception research. It is important to know what subjects believe they are seeing, and subjective reports shed light on this phenomenological question. Such reports may also, however, be affected by biases or demand characteristics. If the results of Experiments 1 and 2 reflect perception and representation of smooth virtual contours and corners under various conditions, it should be possible to find some objective performance task in which these percepts or representations obtained from perception make participants better or worse in a situation where there is an objectively correct answer (c.f., Kellman et al., 2005). We assessed differences in the degree to which dots sampled from smooth and cornered contours were perceived as shapes using objective measures in Experiments 3 and 4.

Experiment 3

One of the key functions that encoding an abstract shape representation serves is allowing comparison of shapes across different orientations (Baker and Kellman, 2018). In Experiment 3, we compared subjects' ability to encode a shape representation for dots sampled from smooth and cornered contours by testing them on a shape matching mental rotation task. Inspired by Shepard and Metzler (1971), we simultaneously presented two differently oriented dot arrays and asked subjects to judge whether they defined the same shape. We expected that if dots sampled from smooth contours are more naturally perceived as shapes, subjects should have an advantage in the mental rotation task on trials where the shape is perceived as smooth.

Methods

Participants

Participants included 25 undergraduates (4 male, 21 female, $M_{age} = 19.8$) from the University of California, Los Angeles who enrolled in the study for course credit. All participants had normal or corrected-to-normal vision.

Stimuli

Smooth and Corner dot arrays were generated as in Experiment 2. In Experiment 3, each array was a member of a pair with either the same shape or a different shape. When the shape was the same, we used the same virtual contour, but sampled a different set of dots so that local spatial relations between dots could not be used as a cue. When the shapes were different, we generated the second member of the pair by moving one of the control points for the original shape a random distance between 1.93° and 4.11° of visual angle toward or away from the center of the shape. We then randomly selected an adjacent control point to the one we just moved and moved it toward or away from the center such that the total contour length for the new shape was the same as the total contour length for the original shape (see Figure 9 for an example pair). For Corner shapes, we then applied the same set of changes described in Experiment 2 to the new shape. Dot arrays also differed in orientation. In each trial, the second dot array could be rotated 0°, 45°, 90°, 135°, or 180° relative to the first.

Design

The experiment consisted of 200 trials, half of which showed shape pairs sampled from smooth virtual contours, and half of which showed shape pairs sampled from cornered contours. For each of these two conditions, there were 20 trials at each of the five magnitudes of rotation, 10 of which included the same shape, and 10 of which included different shapes.

Procedure

On each trial, two arrays of dots were shown on the screen simultaneously, one centered in the left half of the monitor screen, and one centered in the right half. Subjects were instructed to look at both dot arrays and determine whether the shape defined by each array of dots was the same or different, irrespective of a difference in orientation and the local positions of dots. The two dot arrays remained on the screen until subjects responded. Participants were told that response time was being measured, but that they should emphasize responding correctly over responding quickly. Before beginning the main experiment, subjects completed 12 practice trials to familiarize themselves with the task. Performance in the practice trials was not analyzed. A sample trial for each condition is shown in Figure 10.

Results

Following Shepard and Metzler (1971), we analyzed the reaction time only for trials in which the two shapes were the same and subjects responded correctly. Mean response times for each magnitude of rotation are shown in Figure 11A. A 2 (dot array type) × 5 (magnitude of rotation) repeated measures ANOVA confirmed a significant main effect for the type of shape from which the dots were sampled, $F(1,25) = 5.33$, $p = 0.03$, $\eta^2_{partial} = 0.18$ and a significant main effect for magnitude of rotation, $F(4,100) = 4.51$, $p = 0.002$, $\eta^2_{partial} = 0.15$. A linear regression test found a significant overall effect of magnitude of rotation on reaction time, $F(1,25) = 14.59$, $p < 0.001$, $\eta^2_{partial} = 0.37$. The slope was numerically greater for response time as a function of magnitude of rotation for cornered stimuli ($RT = 154 * Deg_{rotated} + 2056$ msec) than for smooth stimuli ($RT = 38 * Deg_{rotated} + 1{,}262$ msec), although the interaction between condition

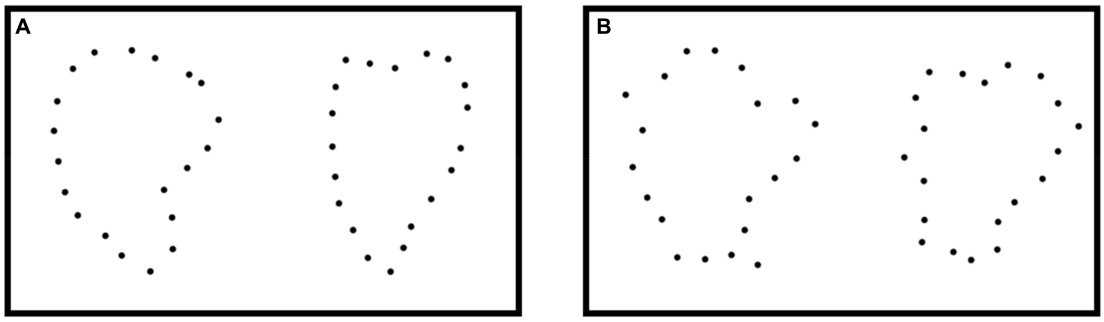


FIGURE 9
Pairs of smooth and angular shapes used in Experiment 3. **(A)** A pair of different shapes from the "Smooth" condition. **(B)** A pair of different shapes from the "Corner" condition.

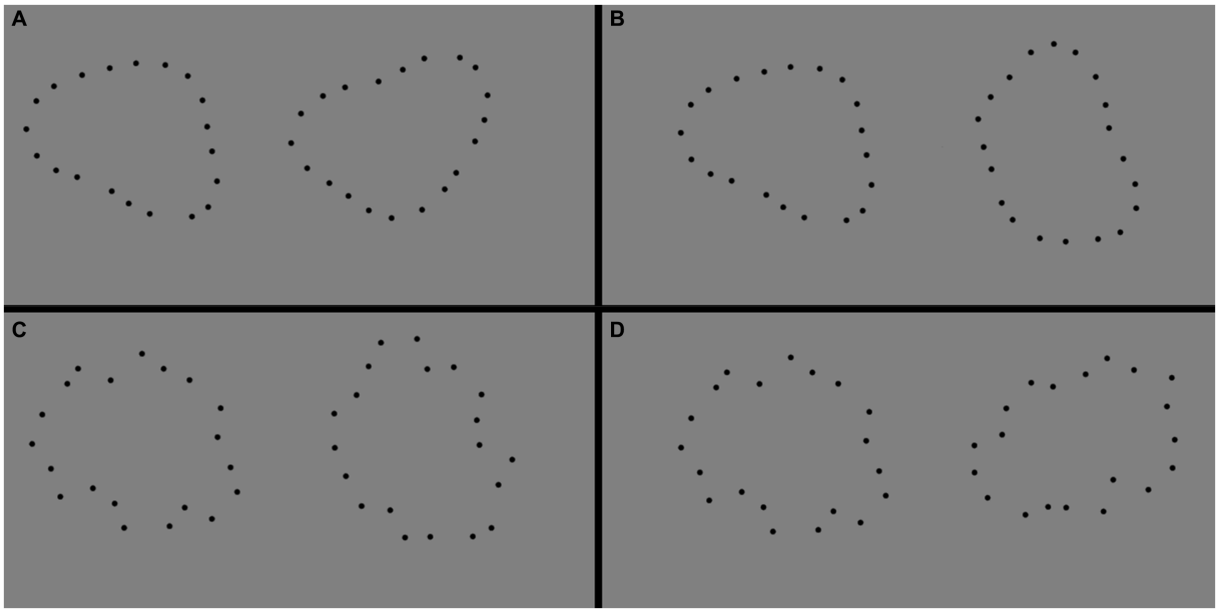


FIGURE 10
Sample trials from Experiment 3. **(A)** "Smooth" trials with the same shape. **(B)** "Smooth" trials with different shape. **(C)** "Corner" contours with the same shape. **(D)** "Corner" contours with different shape with different shape. All shapes are rotated 135°.

and magnitude of rotation was not significant, $F(4,100) = 0.98$, $p = 0.42$, $\eta^2_{partial} = 0.04$.

Superiority in performance for dots sampled from smooth contours was also reflected in sensitivity measures, calculated as the proportion of trials in which participants correctly reported that the shape had changed ("hits") vs. the proportion of trials in which participants incorrectly reported that the shape had changed ("false alarms"; Figure 11B). In trials with no misses or no false alarms, we used the correction recommended by Wickens (2001) of adding half an observation of a miss or false alarm when none was present. A 2 (dot array type) × 5 (magnitude of rotation) repeated measures ANOVA confirmed that sensitivity was significantly higher in displays in which the dots were sampled from smooth contours than displays with dots sampled from contours with sharp corners $F(1,24) = 29.36$, $p < 0.001$, $\eta^2_{partial} = 0.55$. The effect of magnitude of rotation on sensitivity was also significant, $F(4,96) = 4.90$, $p = 0.001$, $\eta^2_{partial} = 0.17$, as was the interaction between dot array type and angle, $F(4,96) = 4.51$, $p = 0.002$, $\eta^2_{partial} = 0.16$.

We also compared participants' bias to report a shape change in each of the 10 conditions. We computed bias as λ_{center}, or the distance between the criterion and the midpoint between the signal and noise distribution. Values of λ_{center} less than 0 indicate a bias to respond "yes" to a shape change and values of λ_{center} greater than 0 indicate a bias to respond "no." The estimates of bias for each condition are plotted in Figure 11C. A repeated measures ANOVA on the estimates of bias found a significant main effect for dot array type on participants' bias, $F(1,24) = 19.36$, $p < 0.001$, $\eta^2_{partial} = 0.45$. Participants were more biased to report a shape change in the smooth condition and more biased to report no shape change in the corner condition. There was also a significant main effect for magnitude of rotation on participants' bias [$F(4,96) = 5.89$, $p < 0.001$, $\eta^2_{partial} = 0.20$]. There was no significant

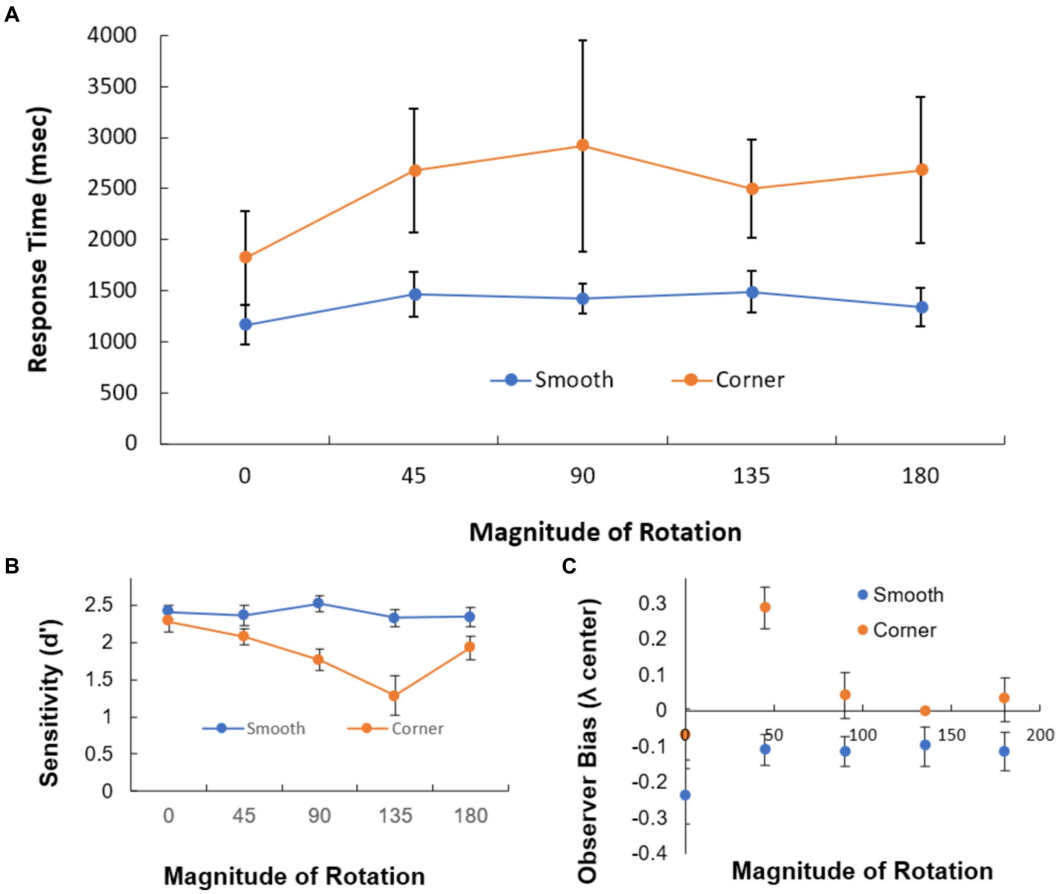


FIGURE 11
Response time, sensitivity, and bias for "Smooth" and "Corner" trials in the mental rotation task of Experiment 3. **(A)** Response time on correct trials as a function of orientation difference between the two displays. **(B)** Sensitivity to shape change. Hits were defined as correct detection of a shape change and false alarms were defined as reports of a shape change when none occurred. **(C)** Response bias in all conditions. λ_{center} reflects the distance between the criterion and the midpoint between the signal and noise distributions. Negative values indicate bias to say there was a shape change. Positive values indicate a bias to say there was not a shape change.

interaction between dot array type and magnitude of rotation on observer bias, $F(4,96) = 2.23$, $p = 0.07$, $\eta^2_{partial} = 0.09$.

Discussion

Experiment 2 found that participants rated dots sampled from contours with cornered contours as less shape-like than dots sampled from smooth curves. In Experiment 3, we tested whether these subjective differences would be reflected in an objective measure of perceptual performance. Because they were rotated and had positions along the contour resampled, the target pairs of dot arrays we showed in Experiment 3 differed from each other both in absolute orientation and in terms of the specific positions of the elements with respect to each other. Accurate responding for the task therefore required forming a representation of a shape's contour from the set of dots that was object-centric and invariant to orientation changes (Baker and Kellman, 2018). Differences in response time and/or sensitivity for the two kinds of dot arrays therefore presumably correspond to the ease with which participants encoded the array as a shape.

We found that dots sampled from shapes with smooth contours could be compared across orientation changes more quickly than dots sampled from shapes with perceived corners, which suggests that these dot arrays are more easily encoded and perceived as orientation-invariant shapes than arrays sampled from shapes with corners. Participants were less accurate when mentally rotating dots sampled from smooth contours than dots sampled from cornered contours. Lower response times therefore cannot be explained by a speed-accuracy tradeoff.

One puzzling aspect of our data is that we found only a small effect of magnitude of orientation difference for the two shapes in each display on response time for either trial type. Slopes in both conditions were flat and explained a smaller proportion of the variance than the Smooth vs. Corner manipulation. The work of Shepard and Metzler (1971), after which we modeled our experiment, showed a strong linear relationship between response time and magnitude of rotation for shapes rotated in the picture plane. Response time has also been shown to vary with degree of change from a canonical orientation in naming tasks for familiar objects (Jolicoeur, 1985). One possibility for why angular difference had such a small effect in our study is that

subjects responded after a somewhat fixed period of time, even if more or less time was needed to make an accurate decision. This could explain why we see a reduction in sensitivity as a function of magnitude of rotation in the sharp corner condition even though response time does not increase. Importantly, though, even if subjects are using a more fixed period of time, this amount of time is different for the smooth and corner conditions. Participants consistently required more time to decide if dot displays sampled from shapes with sharp corners were the same or different, even if response times did not increase monotonically with orientation differences in either condition.

Another intriguing possibility is that dot configurations represent a special class of stimuli whose time for recognition does not scale with magnitude of rotation. Past work on mental rotation has shown that certain kinds of stimuli with salient landmark features have much flatter recognition slopes than stimuli without salient landmarks (Hochberg and Gellman, 1977). Flat slopes have also been found for familiar objects when participants were informed ahead of time what object they would be shown (Cooper and Shepard, 1973). Why mental rotation of dot patterns would have flat slopes is mysterious in view of these findings, as they are neither familiar nor do they have salient local features. In fact, any salient local feature obtained from a local group of dots in one of the arrays would not be present in the other matching array, since dot positions along the contour are independently sampled in matched pairs. One possibility is that the simplicity of dot arrays gives rise to flat mental rotation slopes. According to Hochberg and Gellman (1977), mental rotation of shapes will scale with angular distance if representations must be built up from successive glances. Possibly, the relatively few bits of information in an array of 25 dots can be extracted with only one glance. This is partially supported by previous findings that the spatial positions of an array of 25 dots are registered within the first 30 ms of exposure (Baker and Kellman, 2018).

Experiment 4

As we have discussed, the visual system has a remarkable capacity to form contour representations from unconnected dots. In Experiment 4, we further tested these capabilities by showing dot displays embedded among a field of random noise dots. The experimental paradigm was similar to one devised by Uttal (1973) for dots along a curved or straight line segment. Uttal found that for these simple segments, participants had significantly more trouble detecting the target when it deviated more from a straight line, but there seemed to be little difference for angular vs. curvilinear deviations. In the present study, we tested participants' ability to detect whole forms defined by dots.

To do this, we used a two-interval forced choice (2IFC) paradigm in which one stimulus contained a shape embedded in noise and the other stimulus contained noise alone. Participants' task was to choose the interval that contained a coherent shape. In order to group together and detect the shape of a set of dots in noise, subjects would have to first use some spatial relationships between the dots in the array to identify which dots belonged to a shape outline and which were random. Typically, important cues such as proximity could be potentially misleading for this kind of display. Manipulating the kind of shape contour that the target dots were sampled from, we tested participants' ability to decide which of the two intervals contained a shape and which consisted only of noise dots. We predicted that unlike simple segments, dot arrays sampled from whole shapes with smooth contours would be more easily detected than dots sampled from whole shapes with sharp corners.

Methods

Participants

Twenty-six undergraduates (6 male, 20 female, $M_{age} = 21.6$) from the University of California, Los Angeles participated in this study for course credit. All participants had normal or corrected-to-normal vision. One subject's data was excluded prior to analyzing his results because he did not appear to understand the instructions by the time he had finished the practice portion of the experiment.

Stimuli

Dot arrays from smooth and cornered shape contours were generated as in Experiment 1. In Experiment 4, however, the dots sampled from contours were hidden among 25 distractor dots. Distractor dots were created by uniformly sampling from the rectangular area that contained the target dots. Each trial also included a dot display with no shape. Rather than placing all 50 dots in the other display completely randomly, we created random displays of 25 dots as in Experiment 1, with the only difference being that we moved each dot twice the average distance between dots. This was to create displays with no shape that still had some emptiness in the middle of the array to prevent participants from using that as a low-level cue. We then added 25 dots by uniformly sampling from the encompassing rectangle as in the target displays. Figure 12 shows a target display with dots from a smooth contour, a target display with dots from a corner contour, and a non-target display.

Design

The experiment had two conditions, a Smooth condition, in which the target dots were sampled from a smooth contour, and a Corner condition, in which the target dots were sampled from a corner contour. For both conditions, the target display was shown first in half of the trials second in the other half. There were 120 total trials for each condition. Participants completed 12 practice trials before beginning the main experiment.

Procedure

We used a 2IFC task in which one display consisted of dots sampled from a smooth or cornered contour among noise dots and the other display consisted only of noise dots. Before beginning the experiment, subjects were told they would be looking for shapes hidden in dots. We showed participants 20 (10 Smooth, 10 Corner) examples of the kind of targets they would be asked to detect in the main experiment. The example targets were shown without distractors.

In each trial, we first presented a fixation cross at the center of the screen for 600 ms, then showed the first of the two dot displays for 800 ms. The dot display was then masked by a pattern of black and white dots for 500 ms, after which the second dot display was shown, also for 800 ms. This display was masked for 500 ms, and then subjects were asked to report whether a shape was hidden in the first or second of the two dot displays. Subjects were not cued to look for any specific

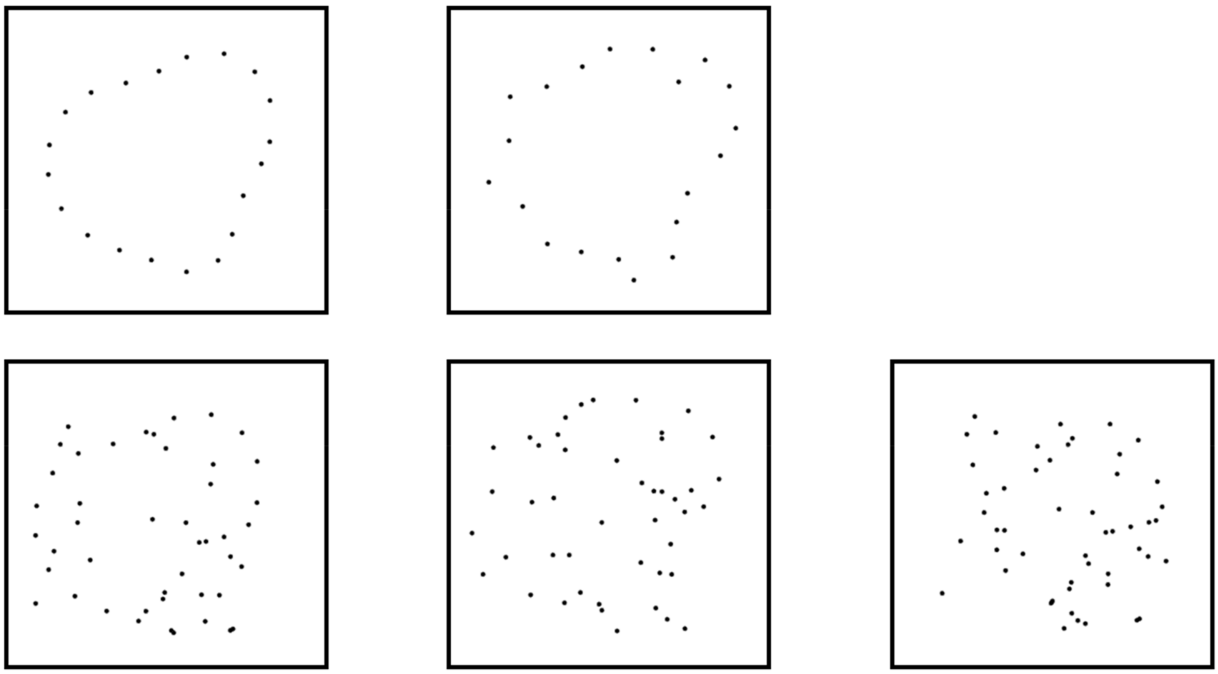


FIGURE 12
Target and distractor displays for Experiment 3. The leftmost column shows a smooth target without (top) and with (bottom) distractor dots added. The middle column shows a cornered target without (top) and with (bottom) distractor dots added. The rightmost column shows a non-target display. In each trial, one of the two kinds of target display (with distractor dots) and the distractor display were presented in a randomized order.

shape in the displays and were told to pick whichever one they thought had dot arrangements that contained any shape. During practice, subjects were given feedback telling them if they were correct or incorrect and showing the hidden shape highlighted in white dots. No feedback was given during the main experiment. A sample trial from the Smooth and Corner condition are shown in Figure 13.

Results

The primary results for Experiment 4 are shown in Figure 14. Performance was significantly better than chance both when the dots were sampled from a smooth contour [$t(24)$ 17.59, $p < 0.001$] and when they were sampled from an cornered contour [$t(24) = 15.69$, $p < 0.001$]. Participants were significantly better at detecting the target shape when the dots were sampled from a smooth contour than when they were sampled from a corner contour [$t(24) = 10.3$, $p < 0.001$, *Cohen's* $d = 2.05$].

Discussion

The results of Experiment 4 show that dot displays embedded in noise were more detectable as shapes when they had been sampled form smooth contours than from cornered ones. In turn, these designations ("smooth" vs. "cornered") were derived from perceptual responses in Experiment 1 (and in prior work by other investigators) to simple dot arrays, consisting of as few as three elements. The superior detection of shapes in noise for the Smooth condition here indicates that these more elementary responses to local dot configurations influence perceptual performance in an objective performance task, and they provide evidence that smooth shapes are more readily encoded from sampled dots. What is required to extract a shape in this task? Surely, relations among elements are crucial, but the task is made more challenging by the fact that the dot elements comprising a shape were physically identical to the distractor dots. The local spatial relationships between small groups of target dots are also not different from relationships between groups of distractor dots or groups that are a mix of targets and distractors. We might compare shape or contour detection to the path detection task developed by Field et al. (1993), which is similar in concept. A crucial difference, however, is that in conventional path detection, individual oriented elements are used, either Gabor patches (Field et al., 1993) or line segments (Pettet, 1999; Baker et al., 2021). In these cases, local orientation relationships described as contour relatability (Kellman and Shipley, 1991) or an association field (Field et al., 1993) are the primary drivers of path detection. Local orientation relationships between the dots determine whether the path is detected depending on the relatability of the local elements. The perceptual salience of paths likely depends on a contour-linking process that produces an intermediate representation in the process of contour interpolation (Kellman et al., 2016; Kellman and Fuchser, 2023).

The situation is different in our Experiment 4. Individual circular dots have no orientation from which relatability can be defined. Unlike the targets used by Uttal (1973), the local spatial relationships between target dots are not consistent. Neither the spacing nor the turning angle is the same between nearby dot triplets in our displays. Detection of the target in our task must depend on more global relations among dots. The visual system might be considering multiple possible dot organizations and determining whether they configure

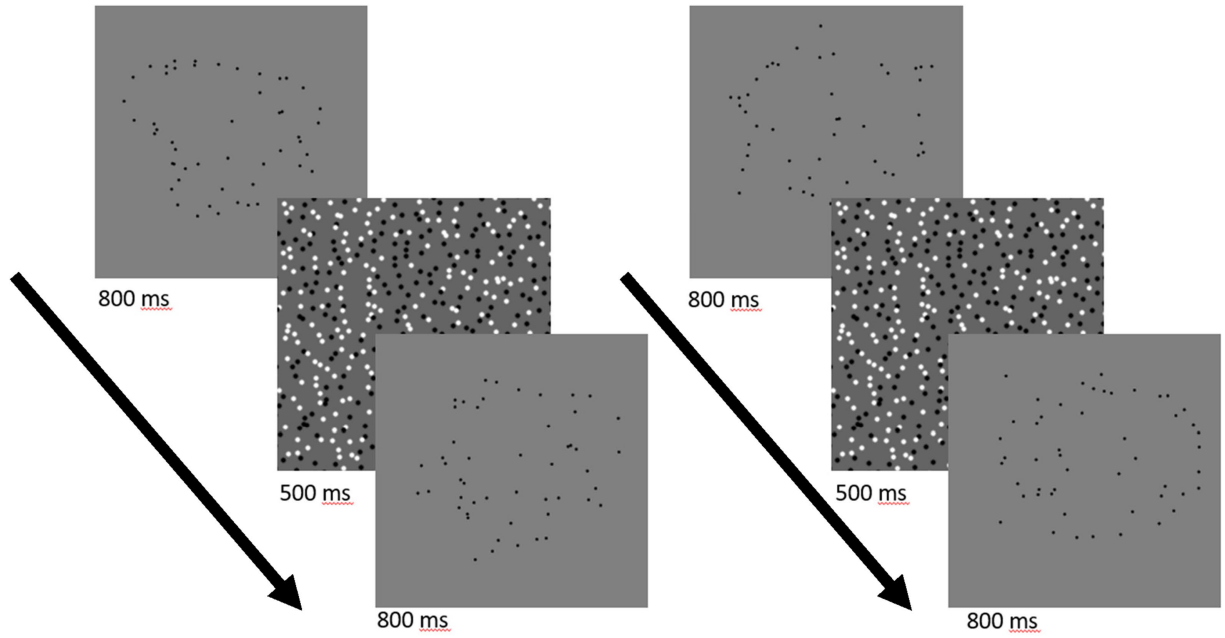


FIGURE 13
Sample trials from Experiment 4. The target shape is sampled from a smooth contour and is shown first (left). The target shape is sampled from an cornered contour and is shown second (right).

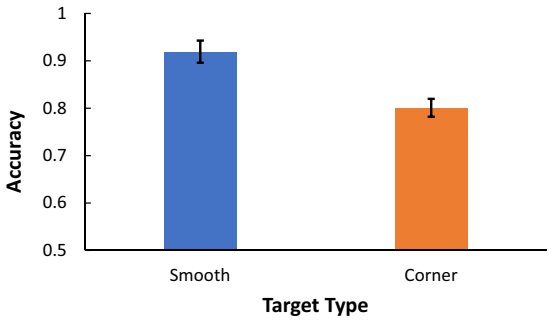


FIGURE 14
Accuracy for detection of shapes in the Smooth and Corner conditions in Experiment 4.

into a global shape. Or, somewhat more local relations among small sets of dots may allow discovery of an extended virtual contour that is perceivable as smooth all along its extent.

Participants' good performance for target dots sampled from both smooth and corner contours suggests that the visual system has a quite robust capability to detect a variety of shapes from distractors. However, there is also a clear performance advantage for detection of shapes with smooth contours over detection of cornered shapes. If detection of perceptually smooth sequences of dots underlies shape detection, dots sampled from smooth shapes exhibit this property along their entire contour, whereas sets of dots perceived as cornered may interrupt perceptual continuity. We develop this idea more fully in the General discussion below.

The perceptual continuity for dots sampled from smooth shapes may also lead to simpler representations than sets of dots sampled from shapes with sharp corners. If first-order discontinuities serve to mark separate parts, or simply indicate important features to be represented, cornered shapes may have more complex representations than smooth ones. Detection of potential connected shapes may be facilitated more by perceptually smooth relations among dots than dot sequences that are more representationally complex in terms of containing corners connecting shorter smooth segments. A similar effect and explanation have been given for search for constant curvature vs. non-constant curvature targets formed by oriented elements (Baker et al., 2021).

General discussion

Perception of contours and shapes from arrangements of separated dots or other tokens is commonplace yet remarkable. Our overall goal in this research was to understand how spatial relations among dots create perceived contours and shapes, and to connect this understanding to general processes of shape perception and representation. Perception of shape from dots offers a special window into abstract shape representations in general. Because no continuous contour is physically present in a display consisting of separated dots, evidence from perceptual tasks that implicates connection, continuity, smoothness, or shape in perceptual representations reveals contributions of processes not directly attributable to the physical stimulus.

In the experiments reported here, we first tested perception of smooth connections vs. corners in small arrays of dots—triplets, or triplets with two additional flanking dots. The results of Experiment 1 indicated that for triplets alone and those with flanking dots that created concave arms, angles relating the triplet dots below 150°–160° more often produced "corner" responses, and "smooth" responses

were very rare for angles less than 100°. With flanking dots that continued the convexity of a triplet, smoothness responses were preserved through a greater angular range, with parity between "smooth" and "corner" responses occurring at about 78°.

Based on these findings of smoothness vs. corner perception determined with small arrays and local angular relations, we tested in further experiments the impact of local relations that were smooth (with triplet angles averaging 154°) vs. cornered (78°–90°) in more complex arrays. In Experiment 2 we compared participants' perception of 25-dot arrays sampled from angular contours with dot arrays sampled from smooth contours. Participants' subjective ratings revealed that dots sampled from smooth shapes were more often and more strongly perceived as coherent shapes than dots sampled from angular shapes. In Experiment 3, we used an objective performance paradigm to assess the effects of local dot relations in processing forms; participants judged as same or different two dot arrays that differed in orientation. We hypothesized that this task would be done, as in classic experiments, by participants mentally rotating one an array to match the orientation of the other, with the expectation that comparison of two dot arrays in this manner would be easier for dot arrangements that were more easily encoded as shapes. We found that participants judged that two dot arrays were the same more quickly and more accurately when they were sampled from smooth contours. In Experiment 4, we embedded a target arrangement of dots defining a virtual contour among an equal number of distractor dots. We found that subjects were more able to detect smooth virtual contours than angular virtual contours, likely because the shape representation the dots give rise to is simpler and therefore easier to search (see Baker et al., 2021 for a similar paradigm).

The results of both Experiments 3 and 4 extend previous research into the perception of dot arrays with a very small number of dots (Pomerantz and Portillo, 2011; Hawkins et al., 2016; Costa and Wagemans, 2021). In those experiments, the configural superiority effect (CSE) paradigm was used to show that an odd-one-out task could be facilitated by the addition of identical elements provided those elements resulted in different emergent features like orientation or proximity in the target display than the distractors. Though the stimuli we used consisted of many more elements than the CSE displays, curvilinear displays were perceived as more configural than displays that contained perceived corners. Experiments 3 and 4 also introduce two additional experimental paradigms, mental rotation and object detection, that can be used as objective indices of the strength of configural structure of dot arrays. The CSE task works extremely well for arrays with a small number of elements to test the effect of local relations between dots. The rotation and detection tasks we used would not work for such sparse displays but showed robust effects for differences in the perception of shape defined by a larger set of elements.

Other research into the organization of dot elements based on Gestalt cues may also explain the perceptual advantage for dot arrays perceived to be curvilinear. For example, notions of similarity might explain why smooth displays, whose vertices were uniformly perceived as curved are more easily perceived as shapes than corner displays, whose vertices would be inhomogeneous, consisting of both perceived curves and perceived corners. Note, however, that these descriptions of inhomogeneity, while related to certain stimulus properties, refer most directly to outcomes of perception (i.e., properties in perceptual representations). The dots in and of themselves are neither corners nor smooth curves. Certain theories of good continuation also predict that smoothness among adjacent pairs or triplets of dots in a sequence facilitates contour perception. According to these theories (e.g., Feldman, 1997; Pizlo et al., 1997; Lezama et al., 2016; Kelly et al., 2024), continuation would be better in displays with fewer extreme deviations from smooth continuation that comes with the addition of perceived corners, which could result in arrays that are more easily resolved into shapes. Definitions of good continuation that explicitly favor collinearity of dots (e.g., Uttal, 1973; van den Berg, 2006) would make the opposite prediction that continuation would be better in corner displays, which have more collinear dots.

Greater facility in encoding shapes with fewer corners and more curvilinear segments would also not be predicted by many other theories of shape and object perception. Much work in middle and high-level vision emphasizes the importance of junctions and non-accidental properties. Geons in Biederman's (1987) work depend crucially on corners and junctions, for example. Under such a theory, we would expect the visual system to be particularly suited to the detection of corners. Indeed, neurophysiological work points to the importance of corners in early visual areas (Heitger et al., 1998) and angular cusps in V4 (Pasupathy and Connor, 2001). Information theoretical work on contour complexity also predicts that objects with straight edges will be perceptually simpler (Attneave, 1954; Norman et al., 2001; Feldman and Singh, 2005). Structural information theory makes the same prediction, positing straight line connections between dots are more economical than curvilinear arcs because arcs are a continuation of both length and angle, thus requiring two bits of information for every one bit of information required for straight line connections (Smits and Vos, 1987, personal communication with Leeuwenberg).

Why, then, are shapes with smooth contours easier to encode than shapes with sharp corners? As Bertamini et al. (2019) point out, there are several reasons to expect angular contours would be more easily processed. Angular contours are comparatively simple to compute, requiring only linear interpolation between salient key points of high curvature (Bertamini et al., 2013). There may also be evolutionary advantages to registering the shapes of angular contours quickly to assess danger (Bar and Neta, 2006). On the other hand, the evolutionary environment in which our visual system evolved likely had many fewer straight edges and sharp angles than the one in which we currently live. Even today, research on scene statistics has found that many of the contours people process in their daily lives are made up of smooth curves (Chow et al., 2002). An analysis of scene statistics can only take us so far, however. The visual system may have evolved to process smooth contours because there were more objects made from smooth contours in our visual environment, but we must still determine what specific visual mechanisms confer this advantage in perceptual processing.

One possibility is that the primitives from which the visual system builds abstract shape representations more easily describe a shape with smooth contours. Elsewhere, we have hypothesized that shape representations are built up from relatively few smoothly joined segments of constant curvature (Garrigan and Kellman, 2011; Kellman et al., 2013; Baker et al., 2021; Baker and Kellman, 2021). Under this theory, corners (first-order or tangent discontinuities) have two important consequences. One is that the presence of a corner would always require spans on either side to be two segment primitives, whereas smoothly changing curvature could be captured by a single segment, provided that the variation in curvature was sufficiently

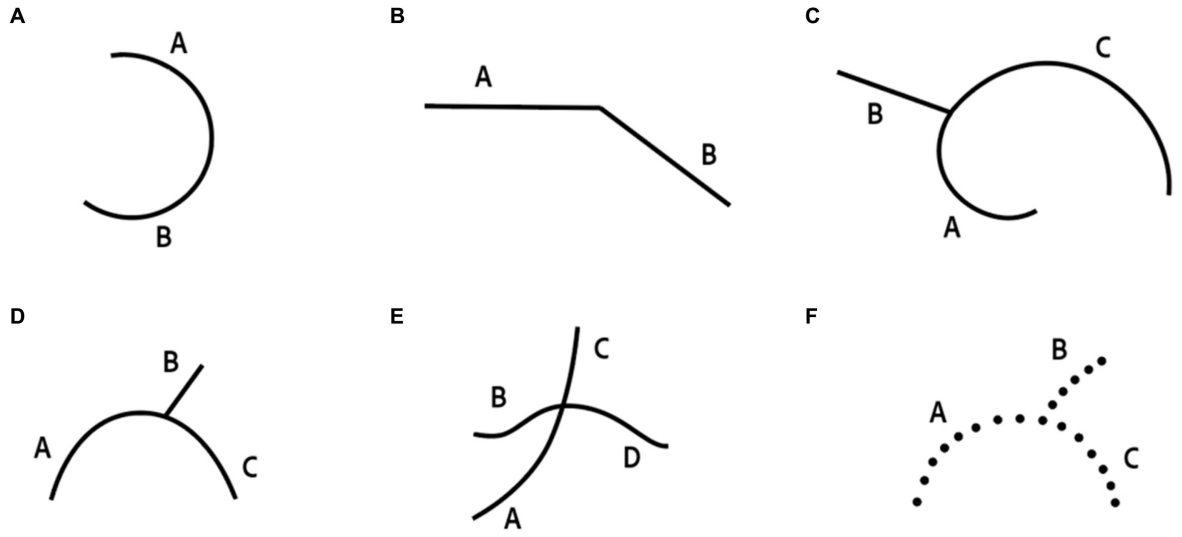


FIGURE 15
Contours made up of more than one curvature segment. **(A)** A contour made up of two smoothly joined constant curvature segments. **(B)** A contour made up of two straight segments joined at a vertex. **(C)** A contour made up of two smoothly joined constant curvature segments and one straight segment joined at a vertex. Individual parts are marked by letters A, B, or C. **(D,E)** Displays show examples redrawn from Wertheimer (1923), corresponding to his Figures 8 and 11, respectively. **(F)** Illustration of the role of perceived corners in dot displays in determining segmentation. Redrawn from Wertheimer (1923), Figure 3.

small. The other consequence is that corners need to be separately encoded in such a representation and may be taken as part boundaries at a certain basic level of representation.

Even when part of a smooth contour requires multiple constant curvature primitives, the smoothly joined segments tend to be perceived as belonging to a single part if they are smooth and monotonic. For example, consider Figures 15A,B. Both displays are made up of two different curvature segments, but the fragment made from two smoothly joined curves looks like a single token, while the fragment made from two straight segments does not.

This observation is closely related to analyses by Wertheimer (1923) in his classic work "Laws of Organization in Perceptual Forms," and it underlies his description of what has come to be called "good continuation." He showed contour displays similar to the ones in Figure 15D and asked observers to divide them into two parts, finding that they almost always organized the two smoothly joined curvature pieces together, separating this connected segment from the straight segment. In terms of derivatives, we may consider any continuous (unbroken) contour to have zero order continuity. Wertheimer's examples show that, despite zero-order continuity, a 1st-order or tangent discontinuity (undefined first derivative) produces some degree of perceptual segmentation. Figure 15E shows another of his examples; here, we can consider 4 segments, A, B, C, and D, and the perceptual impression is that A and C are a unified segment, as are B and D, but observers do not naturally partition such a display into BC and AD, or AB and CD. Although Wertheimer did not invoke presence or absence of discontinuities in the first derivative as the relevant information, he gave a number of examples (in his Figures 1–19), all of which indicate that a contour junction (tangent discontinuity) breaks contiguous line drawings into discernible parts, whereas the smooth continuation (absence of a tangent discontinuity) produces perception of a single contour or contour segment. It is interesting that despite offering two formal names for this principle (the "Factor of Direction" and the "Factor of Good Curve"), it is a phrase he used in passing—"good continuation" that has stuck as the name of this principle.

Figures 15A,C also illustrate that higher-order discontinuities, such as the 2nd-order discontinuity where two curves different smoothly join (matched slope at the join point), do not produce obvious perceptual segmentation. Evidence from visual search in noise shows that search for a contour segment with a 0-order or first-order discontinuity from other segments is easy, but a segment having 0-order and first-order continuity, but a second-order discontinuity, is effortful, slow, and error-prone (Kellman et al., 2003). If shapes made up of sharp corners are perceived to have significantly more parts than shapes made up of smoothly connected contours, it follows that they will be more representationally complex and therefore more difficult to encode.

Consistent with the above reasons that sharp corners may impose an additional encoding burden is that corners are important features for other perceptual processing goals, such as identifying points at which one object might be occluding another (Ratoosh, 1949; Dinnerstein and Wertheimer, 1957; Shipley and Kellman, 1990; Kellman and Shipley, 1991; Rubin, 2001). Also, as we mentioned earlier, corners are important in theories of object representation and recognition (e.g., Biederman, 1987). Although some of these accounts suggest that encoding of corners might be beneficial for comparing objects, it may increase the complexity relative to smooth objects.

Conclusion

The results from these experiments suggest that the visual system perceives shapes from arrays of dots more easily when the perceived

contour between the points is smooth rather than angular. Although consistent with previous literature concerning good continuation between dots, these findings refute other formulations of good continuation that explicitly favor linear continuations. They also point to a more general phenomenon in shape perception that extraction of curvature is a fundamental process in the formation of an abstract shape representation and allows for efficient encoding of contours with changing orientation. Virtual contours that can be described by a relatively constrained set of curvature primitives appear to give rise to shapes more often, more quickly, and more precisely than virtual contours that are perceived and represented as segments connected by corners.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: DOI: 10.17605/OSF.IO/7DGJC.

Ethics statement

The studies involving humans were approved by IRB University of California, Los Angeles. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

NB: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. PK: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. We gratefully acknowledge support from the National Institutes of Health award number R01 CA236791 to PK.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Attneave, F. (1954). Some informational aspects of visual perception. *Psychol. Rev.* 61, 183–193. doi: 10.1037/h0054663

Baker, N., Garrigan, P., and Kellman, P. J. (2021). Constant curvature segments as building blocks of 2D shape representation. *J. Exp. Psychol. Gen.* 150, 1556–1580. doi: 10.1037/xge0001007

Baker, N., and Kellman, P. J. (2018). Abstract shape representation in human visual perception. *J. Exp. Psychol. Gen.* 147, 1295–1308. doi: 10.1037/xge0000409

Baker, N., and Kellman, P. J. (2021). Constant curvature modeling of abstract shape representation. *PLoS One* 16:e0254719. doi: 10.1371/journal.pone.0254719

Bar, M., and Neta, M. (2006). Humans prefer curved visual objects. *Psychol. Sci.* 17, 645–648. doi: 10.1111/j.1467-9280.2006.01759.x

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and brain sciences.* 22, 577–660.

Barsalou, L. W. (2003). Situated simulation in the human conceptual system. *Language and cognitive processes.* 18, 513–562.

Bertamini, M., Helmy, M., and Bates, D. (2013). The visual system prioritizes locations near corners of surfaces (not just locations near a corner). *Atten. Percept. Psychophys.* 75, 1748–1760. doi: 10.3758/s13414-013-0514-1

Bertamini, M., Palumbo, L., and Redies, C. (2019). An advantage for smooth compared with angular contours in the speed of processing shape. *J. Exp. Psychol. Hum. Percept. Perform.* 45, 1304–1318. doi: 10.1037/xhp0000669

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115–147. doi: 10.1037/0033-295X.94.2.115

Bouma, H. (1976). Perceptual functions. In J. A. Michon, E. G. Eijkman and KlerkL. F. De. *Dutch handbook of psychonomy* (pp. 229–287). Deventer, Netherlands: Van Loghem Slaterus (in Dutch).

Box, G. E., and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* 26, 211–243.

Chow, C. C., Jin, D. Z., and Treves, A. (2002). Is the world full of circles? *J. Vis.* 2:4. doi: 10.1167/2.8.4

Clowes, M. B. (1971). On seeing things. *Artif Intell* 2, 79–116. doi: 10.1016/0004-3702(71)90005-1

Cooper, L. A., and Shepard, R. N. (1973). "Chronometric studies of the rotation of mental images" in *Visual information processing*. ed. Chase W. G. (New York: Academic Press), 75–176.

Costa, T. L., and Wagemans, J. (2021). Gestalts at threshold could reveal gestalts as predictions. *Sci. Rep.* 11:18308. doi: 10.1038/s41598-021-97878-0

De Winter, J., and Wagemans, J. (2008). Perceptual saliency of points along the contour of everyday objects: a large-scale study. *Percept. Psychophys.* 70, 50–64. doi: 10.3758/PP.70.1.50

Dinnerstein, D., and Wertheimer, M. (1957). Some determinants of phenomenal overlapping. *Am. J. Psychol.* 70, 21–37. doi: 10.2307/1419226

Feldman, J. (1996). Regularity vs genericity in the perception of collinearity. *Perception* 25, 335–342. doi: 10.1068/p250335

Feldman, J. (1997). Curvilinearity, covariance, and regularity in perceptual groups. *Vision Res.* 37, 2835–2848. doi: 10.1016/S0042-6989(97)00096-5

Feldman, J., and Singh, M. (2005). Information along contours and object boundaries. *Psychol. Rev.* 112, 243–252. doi: 10.1037/0033-295X.112.1.243

Field, D. J., Hayes, A., and Hess, R. F. (1993). Contour integration by the human visual system: evidence for a local "association field". *Vision Res.* 33, 173–193. doi: 10.1016/0042-6989(93)90156-Q

Garrigan, P., and Kellman, P. J. (2011). The role of constant curvature in 2-D contour shape representations. *Perception* 40, 1290–1308. doi: 10.1068/p6970

Hawkins, R. X., Houpt, J. W., Eidels, A., and Townsend, J. T. (2016). Can two dots form a gestalt? Measuring emergent features with the capacity coefficient. *Vision Res.* 126, 19–33. doi: 10.1016/j.visres.2015.04.019

Heitger, F., Von Der Heydt, R., Peterhans, E., Rosenthaler, L., and Kübler, O. (1998). Simulation of neural contour mechanisms: representing anomalous contours. *Image Vis. Comput.* 16, 407–421. doi: 10.1016/S0262-8856(97)00083-8

Hess, R., and Field, D. (1999). Integration of contours: new insights. *Trends Cogn. Sci.* 3, 480–486. doi: 10.1016/S1364-6613(99)01410-2

Hochberg, J., and Gellman, L. (1977). The effect of landmark features on mental rotation times. *Mem. Cognit.* 5, 23–26. doi: 10.3758/BF03209187

Jolicoeur, P. (1985). The time to name disoriented natural objects. *Mem. Cognit.* 13, 289–303. doi: 10.3758/BF03202498

Kalar, D. J., Garrigan, P., Wickens, T. D., Hilger, J. D., and Kellman, P. J. (2010). A unified model of illusory and occluded contour interpolation. *Vision Res.* 50, 284–299. doi: 10.1016/j.visres.2009.10.011

Kanizsa, G. (1979). *Organization in vision: Essays on gestalt perception*. Praeger Press.

Kellman, P., Erlikhman, G., and Carrigan, S. (2016). Is there a common mechanism for path integration and illusory contour formation? *J. Vis.* 16:311. doi: 10.1167/16.12.311

Kellman, P. J., and Fuchser, V. (2023). *Visual completion and intermediate*. Sensory Individuals: Unimodal and Multimodal Perspectives, 55.

Kellman, P. J., and Massey, C. M. (2013). "Perceptual learning, cognition, and expertise," in *The psychology of learning and motivatio.* (Ed.) S. J. Dickinson and Z. Pizlo, Vol. 58, (Amsterdam: Elsevier Inc.), 117–165.

Kellman, P. J., Garrigan, P. B., Kalar, D., and Shipley, T. F. (2003). Good continuation and relatability: related but distinct principles. *J. Vis.* 3:120. doi: 10.1167/3.9.120

Kellman, P. J., Garrigan, P., Shipley, T. F., Yin, C., and Machado, L. (2005). 3-d interpolation in object perception: evidence from an objective performance paradigm. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 558–583. doi: 10.1037/0096-1523.31.3.558

Kellman, P. J., and Shipley, T. F. (1991). A theory of visual interpolation in object perception. *Cogn. Psychol.* 23, 141–221. doi: 10.1016/0010-0285(91)90009-D

Kelly, B. A., Kemp, C., Little, D. R., Hamacher, D., and Cropper, S. J. (2024). Visual perception principles in constellation creation. *Top. Cogn. Sci.* 16, 25–37. doi: 10.1111/tops.12720

Kemp, C., Hamacher, D. W., Little, D. R., and Cropper, S. J. (2022). Perceptual grouping explains similarities in constellations across cultures. *Psychol. Sci.* 33, 354–363. doi: 10.1177/09567976211044157

Koffka, K. (1931). "Psychology of visual perception" in *Handbook of normal and pathological physiology*. ed. A. Bethe (Berlin: Dessoir), 1215–1271.

Koffka, K. "*Principles of gestalt psychology, 481–493*." (Routledge) (1935).

Krzywinski, M., and Altman, N. (2014). Visualizing samples with box plots. *Nat. Methods* 11, 119–120. doi: 10.1038/nmeth.2813

Kubovy, M., and Wagemans, J. (1995). Grouping by proximity and multistability in dot lattices: a quantitative gestalt theory. *Psychol. Sci.* 6, 225–234. doi: 10.1111/j.1467-9280.1995.tb00597.x

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Percept. Psychophys.* 63, 1279–1292. doi: 10.3758/BF03194543

Lezama, J., Randall, G., Morel, J. M., and von Gioi, R. G. (2016). Good continuation in dot patterns: a quantitative approach based on local symmetry and non-accidentalness. *Vision Res.* 126, 183–191. doi: 10.1016/j.visres.2015.09.004

McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of box plots. *Am. Stat.* 32, 12–16. doi: 10.1080/00031305.1978.10479236

Metzger, W. (2009). *Laws of seeing*. Cambridge, MA: Mit Press.

Norman, J. F., Phillips, F., and Ross, H. E. (2001). Information concentration along the boundary contours of naturally shaped solid objects. *Perception* 30, 1285–1294. doi: 10.1068/p3272

O'Callaghan, J. F. (1974). Computing the perceptual boundaries of dot patterns. *Comput Graph Image Process* 3, 141–162. doi: 10.1016/S0146-664X(74)80004-3

Papari, G., and Petkov, N. (2005). "Algorithm that mimics human perceptual grouping of dot patterns" in *International symposium on brain, vision, and artificial intelligence*. (Eds.) Gregorio, M. D., Frucci, D. M., and Musio, C. (Berlin, Heidelberg: Springer), 497–506.

Pasupathy, A., and Connor, C. E. (2001). Shape representation in area V4: position-specific tuning for boundary conformation. *J. Neurophysiol.* 86, 2505–2519. doi: 10.1152/jn.2001.86.5.2505

Pettet, M. W. (1999). Shape and contour detection. *Vision Res.* 39, 551–557. doi: 10.1016/S0042-6989(98)00130-8

Pizlo, Z., Salach-Golyska, M., and Rosenfeld, A. (1997). Curve detection in a noisy image. *Vision Res.* 37, 1217–1241. doi: 10.1016/S0042-6989(96)00220-9

Pomerantz, J. R., and Portillo, M. C. (2011). Grouping and emergent features in vision: toward a theory of basic gestalts. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 1331–1349. doi: 10.1037/a0024330

Prinzmetal, W., and Banks, W. P. (1977). Good continuation affects visual detection. *Percept. Psychophys.* 21, 389–395. doi: 10.3758/BF03199491

Ratoosh, P. (1949). On interposition as a cue for the perception of distance. *Proc. Natl. Acad. Sci. U. S. A.* 35, 257–259. doi: 10.1073/pnas.35.5.257

Rubin, N. (2001). The role of junctions in surface completion and contour matching. *Perception* 30, 339–366. doi: 10.1068/p3173

Shepard, R. N., and Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science* 171, 701–703. doi: 10.1126/science.171.3972.701

Shipley, T. F., and Kellman, P. J. (1990). The role of discontinuities in the perception of subjective figures. *Percept. Psychophys.* 48, 259–270. doi: 10.3758/BF03211526

Smits, J. T. S., and Vos, P. G. (1986). A model for the perception of curves in dot figures: the role of local salience of "virtual lines". *Biol. Cybern.* 54, 407–416. doi: 10.1007/BF00355546

Smits, J. T., and Vos, P. G. (1987). The perception of continuous curves in dot stimuli. *Perception* 16, 121–131. doi: 10.1068/p160121

Smits, J. T., Vos, P. G., and Van Oeffelen, M. P. (1985). The perception of a dotted line in noise: a model of good continuation and some experimental results. *Spat. Vis.* 1, 163–177. doi: 10.1163/156856885X00170

Troncoso, X. G., Macknik, S. L., and Martinez-Conde, S. (2005). Novel visual illusions related to Vasarely's 'nested squares' show that corner salience varies with corner angle. *Perception* 34, 409–420. doi: 10.1068/p5383

Uttal, W. R. (1973). The effect of deviations from linearity on the detection of dotted line patterns. *Vision Res.* 13, 2155–2163. doi: 10.1016/0042-6989(73)90193-4

Van Assen, M. A., and Vos, P. G. (1999). Evidence for curvilinear interpolation from dot alignment judgements. *Vision Res.* 39, 4378–4392. doi: 10.1016/S0042-6989(99)00150-9

van den Berg, M. (2006). *Grouping by proximity and grouping by good continuation in the perceptual organization of random dot patterns*. Wickens: New York, NY: Unpublished doctoral dissertation, University of Virginia, Charlottesville.

Van Oeffelen, M. P., and Vos, P. G. (1983). An algorithm for pattern description on the level of relative proximity. *Pattern Recogn.* 16, 341–348. doi: 10.1016/0031-3203(83)90040-7

Wertheimer, M. (1923). "Laws of organization in perceptual forms" in *A source book of Gestalt Psychology*. (Ed.) W. D. Ellis, London: Kegan Paul, Trench, Trubner & Co.

Wickens, T. D. (2001). *Elementary signal detection theory* Oxford university press.

Wouterlood, D., and Boselie, F. (1992). A good-continuation model of some occlusion phenomena. *Psychol. Res.* 54, 267–277. doi: 10.1007/BF01358264

Yuen, H. K., Princen, J., Illingworth, J., and Kittler, J. (1990). Comparative study of Hough transform methods for circle finding. *Image Vis. Comput.* 8, 71–77. doi: 10.1016/0262-8856(90)90059-E

Zucker, S. W., Stevens, K. A., and Sander, P. (1983). The relation between proximity and brightness similarity in dot patterns. *Percept. Psychophys.* 34, 513–522. doi: 10.3758/BF03205904

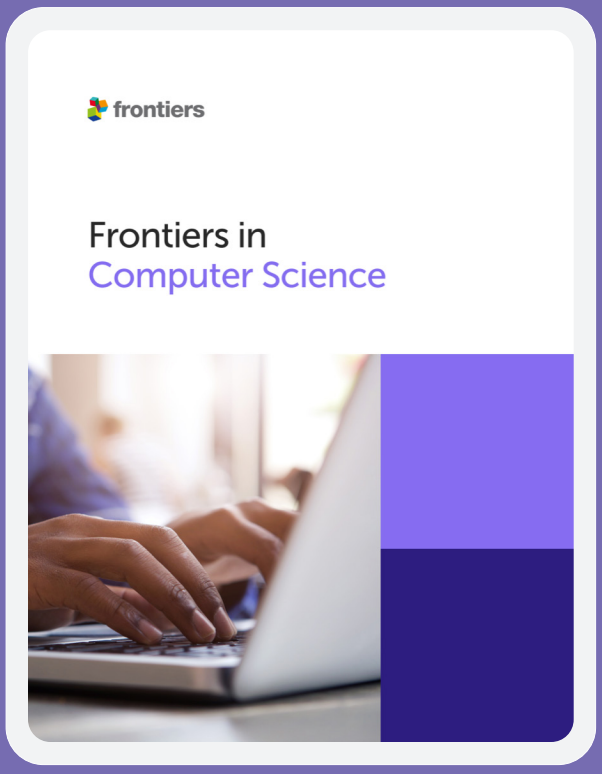
Frontiers in Computer Science

Explores fundamental and applied computer science to advance our understanding of the digital era

An innovative journal that fosters interdisciplinary research within computational sciences and explores the application of computer science in other research domains.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact