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

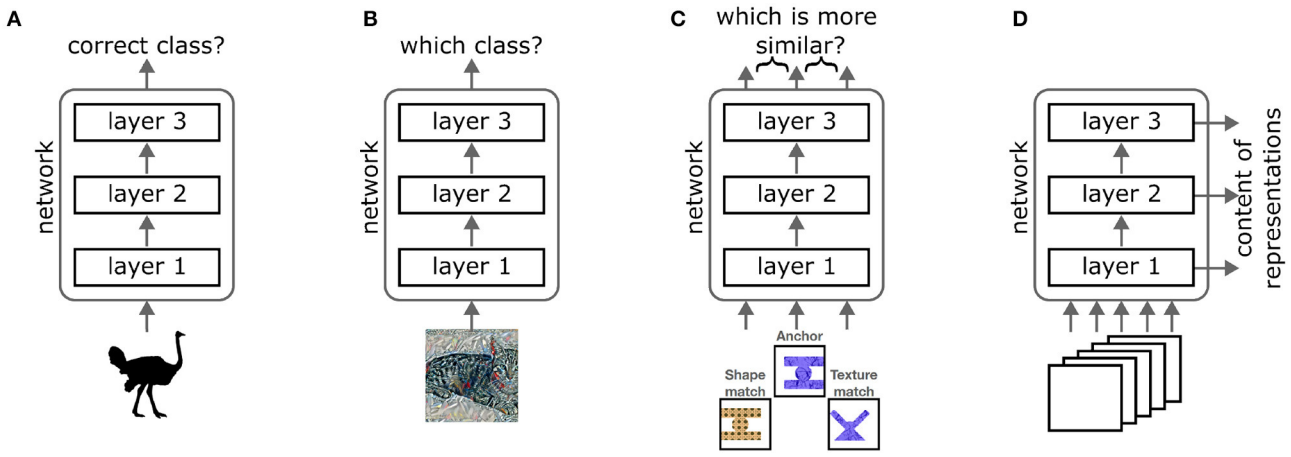

FIGURE 1
The different types of experiments used to assess shape processing in neural networks. **(A)** Diagnostic stimuli restrict the amount of information available in an image. For example, silhouettes only show the shape of an object. If a network can classify such stimuli correctly, it must be 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).

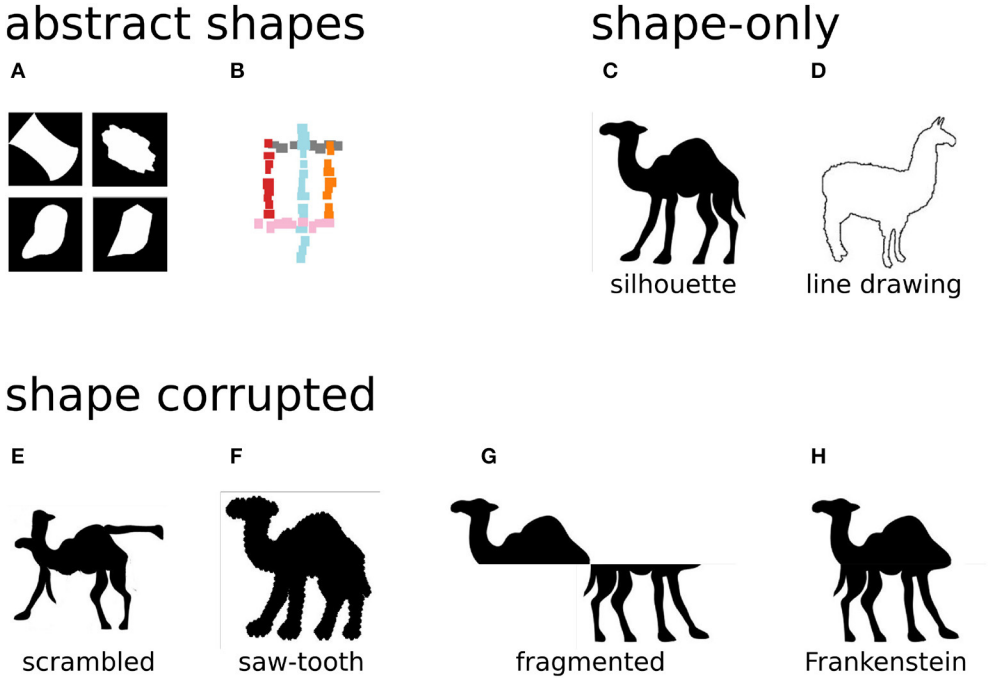

FIGURE 2
Diagnostic stimuli isolate or manipulate shape information. Artificial, abstract shapes allow tight control over which shape information is present: **(A)** Kalfas et al. (2018) generated shapes in four categories: regular (top left), complex (top right), simple curved (bottom left), and simple 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.

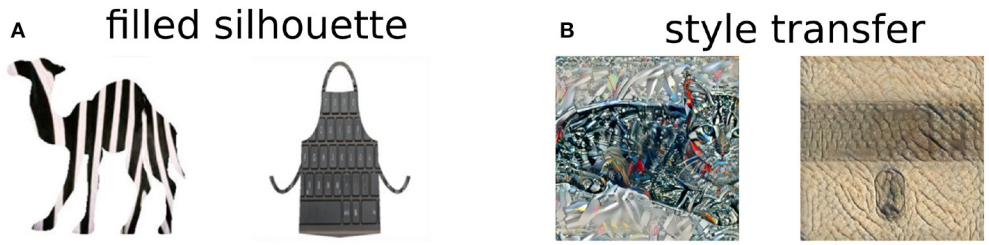

FIGURE 3
Stimuli used in cue conflict experiments. **(A)** Silhouettes filled with surface texture of another class. **(B)** Cue conflict generated using style transfer, using an image of one class as content or shape and an image of another class as style. The left image has shape "cat" and texture "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.

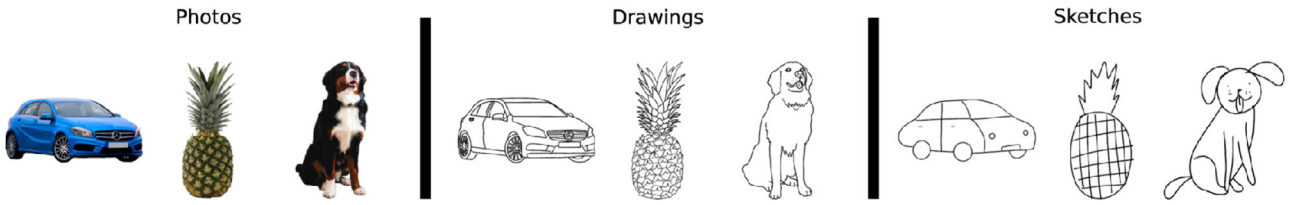

FIGURE 4
Stimuli used in Singer et al. (2022). Line drawings closely match the photographs, such that the represented shape information is similar while the surface properties differ. Line drawings and sketches have similar (absent) surface properties, but show different degrees of shape information.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

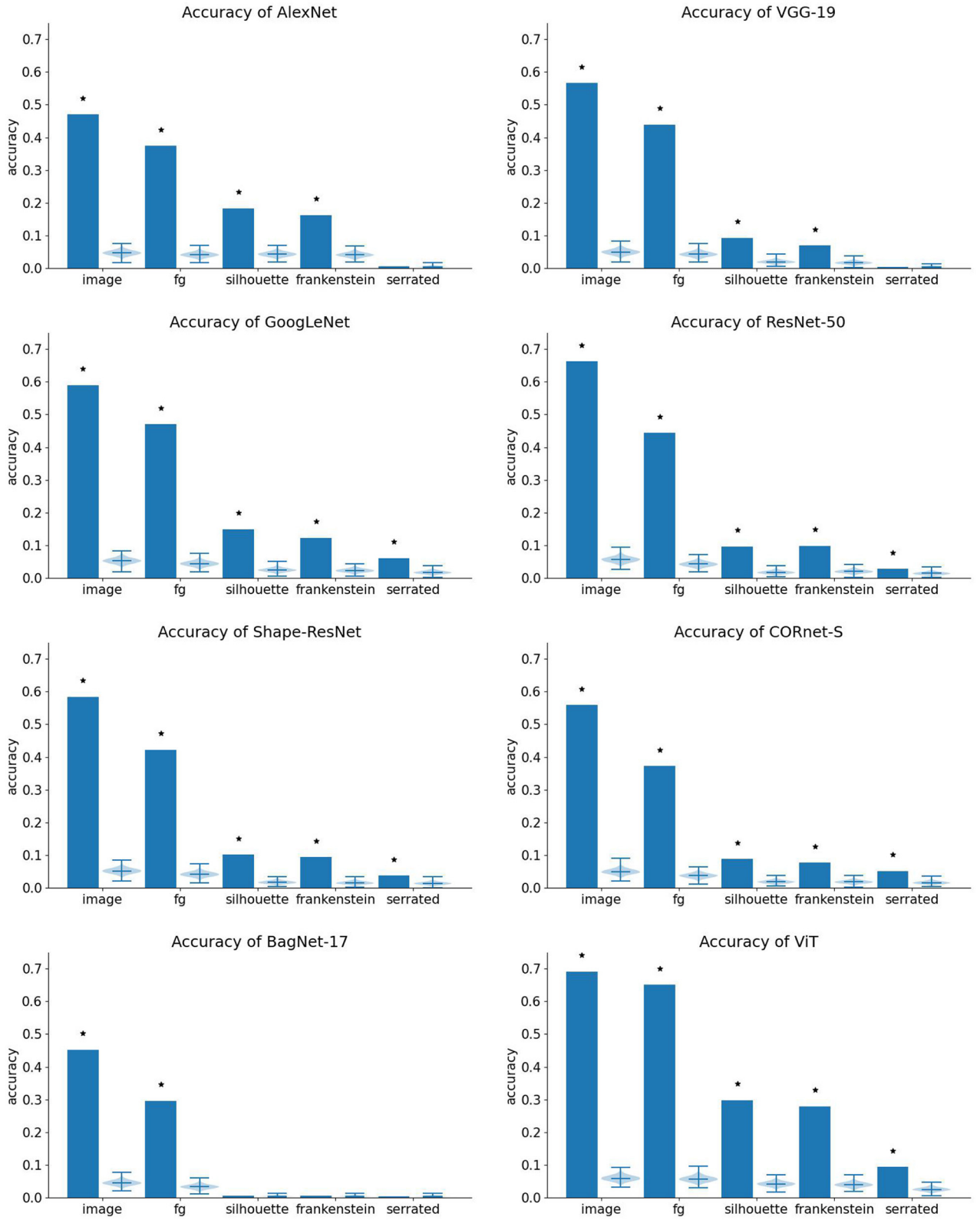

FIGURE 6
Classification accuracy of each network. Bar height indicates performance on the respective dataset. The violin plot to the right of each bar shows the corresponding null distribution. Stars indicate that accuracy is significantly higher than chance.

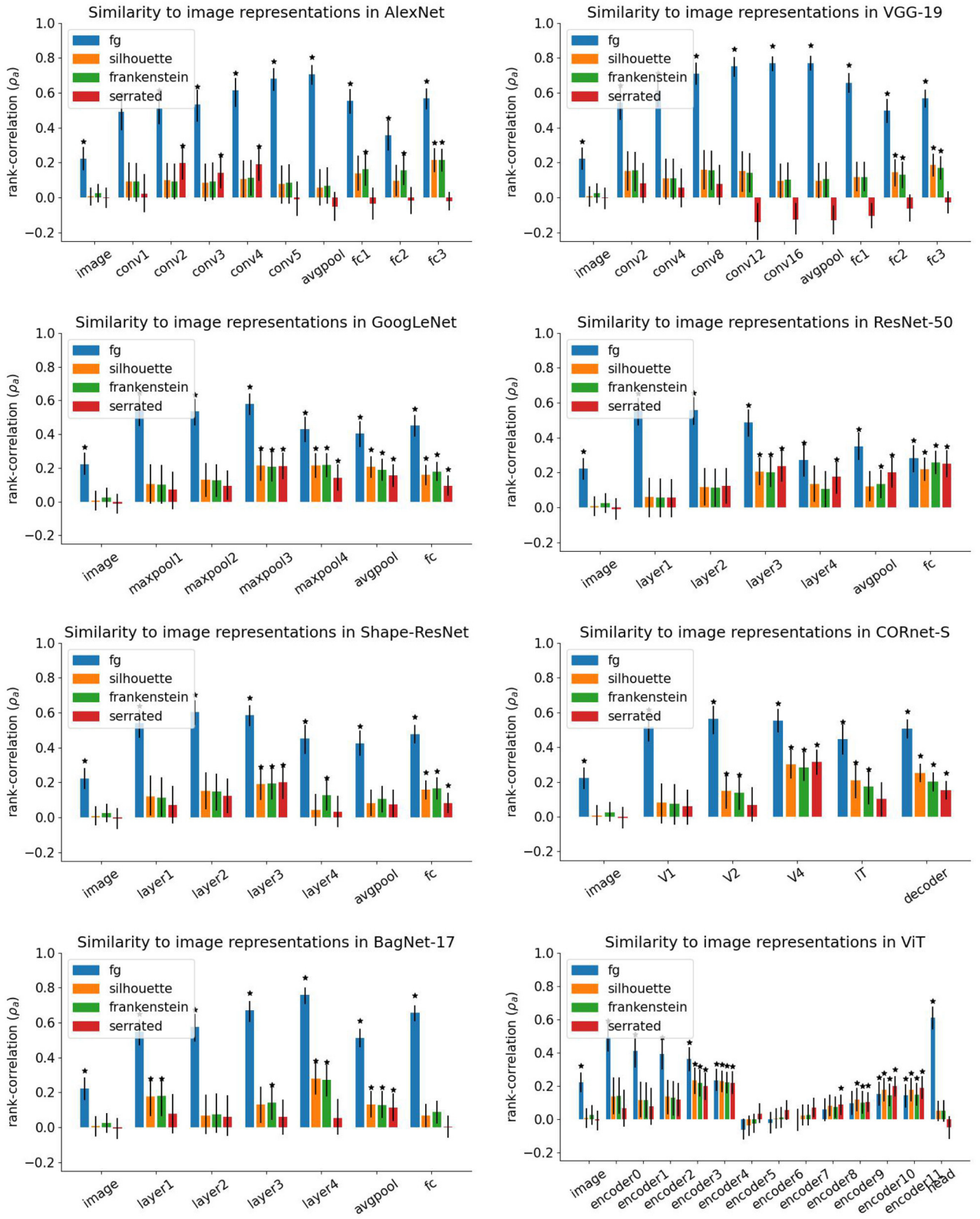

FIGURE 7
Results of representational similarity analysis. Height of bars indicates average rank-correlation over bootstrap runs. Error bars indicate 95% confidence intervals from bootstrap runs. Stars mark similarities that are significantly larger than 0.

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