frontiers in Computational Neuroscience

HYPOTHESIS AND THEORY
published: 25 March 2022
doi: 10.3389/fncom.2022.857653

Maximal Dependence Capturing as a Principle of Sensory Processing

*Rishabh Raj[1], Dar Dahlen[1], Kyle Duyck[1] and C. Ron Yu[1,2]**

[1] Stowers Institute for Medical Research, Kansas City, MO, United States, [2] Department of Anatomy and Cell Biology, University of Kansas Medical Center, Kansas City, KS, United States

Sensory inputs conveying information about the environment are often noisy and incomplete, yet the brain can achieve remarkable consistency in recognizing objects. Presumably, transforming the varying input patterns into invariant object representations is pivotal for this cognitive robustness. In the classic hierarchical representation framework, early stages of sensory processing utilize independent components of environmental stimuli to ensure efficient information transmission. Representations in subsequent stages are based on increasingly complex receptive fields along a hierarchical network. This framework accurately captures the input structures; however, it is challenging to achieve invariance in representing different appearances of objects. Here we assess theoretical and experimental inconsistencies of the current framework. In its place, we propose that individual neurons encode objects by following the principle of maximal dependence capturing (MDC), which compels each neuron to capture the structural components that contain maximal information about specific objects. We implement the proposition in a computational framework incorporating dimension expansion and sparse coding, which achieves consistent representations of object identities under occlusion, corruption, or high noise conditions. The framework neither requires learning the corrupted forms nor comprises deep network layers. Moreover, it explains various receptive field properties of neurons. Thus, MDC provides a unifying principle for sensory processing.

Keywords: object recognition (OR), computational modeling, invariant representation, sparse recovery (SR), redundancy reduction, redundancy capturing, sparse coding, grandmother cell

OPEN ACCESS

Edited by:
Paul Miller,
Brandeis University, United States

Reviewed by:
Tony Lindeberg,
Royal Institute of Technology, Sweden
Gianluca Serafini,
San Martino Hospital (IRCCS), Italy

****Correspondence:***
C. Ron Yu
cry@stowers.org

Received: *18 January 2022*
Accepted: *15 February 2022*
Published: *25 March 2022*

Citation:
Raj R, Dahlen D, Duyck K and Yu CR (2022) Maximal Dependence Capturing as a Principle of Sensory Processing. Front. Comput. Neurosci. 16:857653. doi: 10.3389/fncom.2022.857653

INTRODUCTION

The world is organized into objects that form the basis of our daily experience. Objects can be assigned meanings from their associations with others and can predict future events. Object recognition is a subject of intensive study in neuroscience, machine vision, and artificial intelligence. It refers to a collection of problems that involve identifying an object from varying input patterns. The most studied object recognition tasks include image segmentation, size and location invariance, representation of 3-D images, identifying occluded objects, and object classification.

Object recognition is a hard problem because a given object can turn up in numerous appearances, can be obscured and occluded, yet the brain readily recognizes it (Ullman, 1996; Riesenhuber and Poggio, 1999b; DiCarlo and Cox, 2007). The prevailing framework of object recognition divides the problem into two subproblems: representation and decision (DiCarlo and Cox, 2007). Objects are thought to be represented in the cortical regions, which are then distinguished and properly classified. With this division are two sets of difficulties. The first is to identify the computational rules that allow the transformation of sensory inputs into accurate and invariant representations. The canonical framework is vision-centric, based on our understanding of highly advanced visual systems such as those in primates. Under this framework, object representation is achieved through hierarchically organized networks of neurons (Riesenhuber and Poggio, 2000; Serre, 2014). In the meantime, it is also premised on efficient coding, where individual neurons transmit information efficiently. There are no intrinsic computation rules that allow the mapping of variant inputs onto the same representation. The second set of difficulties is related to resolving the ambiguities in the representations and determining whether a representation belongs to a specific object or a class of objects. The canonical solution to this problem is for a network to be trained using a large set of examples to achieve statistical robustness in object identification. How different representations are properly classified based on statistical learning has created difficult challenges (Chen et al., 2018; Fiser and Aslin, 2001; Turk-Browne et al., 2005; DiCarlo and Cox, 2007).

Animals appear to solve object recognition seamlessly. They exhibit the acute ability to discriminate similar sensory inputs and identify objects in complex natural environments to guide their behaviors, often in a fraction of a second. Species with limited brain complexity can perform robust recognition. Juxtaposing the ease of most animals' seemingly effortless ability to recognize objects and the difficulty of current models to solve this problem, one must ask, *why is object recognition hard*? We suggest that the vision-centric approaches have largely ignored the ethological need for object recognition from the perspective of animals. They do not provide the explanatory power to other senses, nor to visual systems that are less sophisticated but equally powerful in performing object recognition. The models may have incorporated specific elements from the visual system that are not needed for object recognition in general and created unnecessary complexity and difficulties in the field. Auditory and olfactory objects are recognizable entities with direct ethological relevance, but their input patterns are less defined than visual input. The neural circuits that process auditory or olfactory information do not have deep structures, but the recognition of odor or audio objects is nonetheless robust. Visual recognition is also strong even in species with simpler and less organized visual systems, such as rodents, arachnids, and insects. A general theory of object recognition must accommodate these systems.

In this article, we assess the assumptions, the framing, and key concepts in the current framework and offer new perspectives of the problem. We introduce an information-theoretical definition of object recognition and hypothesize that maximal dependence capturing is a general principle in sensory processing. We present evidence from mathematical simulations that the new framework allows robust object representation. At the same time, it also provides the power to interpret the firmly established experimental observations that form the basis of current models.

A CRITIQUE OF THE CURRENT FRAMEWORK OF OBJECT RECOGNITION

The classic framing of the object recognition problem has been representational, i.e., neurons faithfully represent sensory features (Hubel and Wiesel, 1962, 1968). Object images are decomposed into elemental components that are processed across various stages (**Figure 1A**). This parts-based decomposition was the idea behind the computations performed in perceptron (Rosenblatt, 1957, 1958) and the theory of *recognition by components* (Biederman, 1985, 1987). Guided by brain anatomy and physiology, later theories propose the hierarchical assembly of elemental features into increasingly complex structures across multiple stages of sensory processing (Barlow, 1961; Atick, 1992; Riesenhuber and Poggio, 1999b; Ullman et al., 2002). In this set of theories, combinations of largely independent features form the basis of brain representations of the objects (Hubel and Wiesel, 1962; Marr, 1969; Marr and Nishihara, 1978; DiCarlo et al., 2012). The framework successfully explains the increase in size and complexity of receptive fields in the mammalian brains (Boussaoud et al., 1991; Rolls and Milward, 2000; Rolls, 2001). It can attain shift and scale invariance while representing 2-dimensional images (Fukushima and Miyaek, 1982; Anderson and Vanessen, 1987; Olshausen et al., 1993). Object representations can also be robust against occlusion (Fukushima, 2003, 2005; Johnson and Olshausen, 2005).

In more recent studies, numerous layers of convolutional and recurrent neural networks are trained to perform specific tasks (Yamins and DiCarlo, 2016a; Richards et al., 2019). These deep learning models can perform at levels that rival or exceed human performances (Cireşan D. C. et al., 2011; Cireşan D. et al., 2011; Sermanet and LeCun, 2011; Krizhevsky et al., 2012; Russakovsky et al., 2015). The success in deep learning and other AI approaches have reinforced the notion that the hierarchical architecture and computational rules associated with it may just be what neuroscientists have been looking for. Indeed, recent efforts have been comparing deep neural networks (DNNs) with brain structures in performing certain tasks (Lee et al., 2007; Yamins and DiCarlo, 2016b; Ponce et al., 2019; Bao et al., 2020).

However, building something to perform a similar function does not mean we have reproduced biology – an airplane uses completely different mechanics from birds to fly. Questions arise whether the DNNs recapitulate the inner working of the brain and if there is a need to engineer less artificial intelligence (Sinz et al., 2019). We believe that there are fundamental conceptual problems inherent in the current framework. Here, we wish to identify these problems. We do not clearly distinguish between

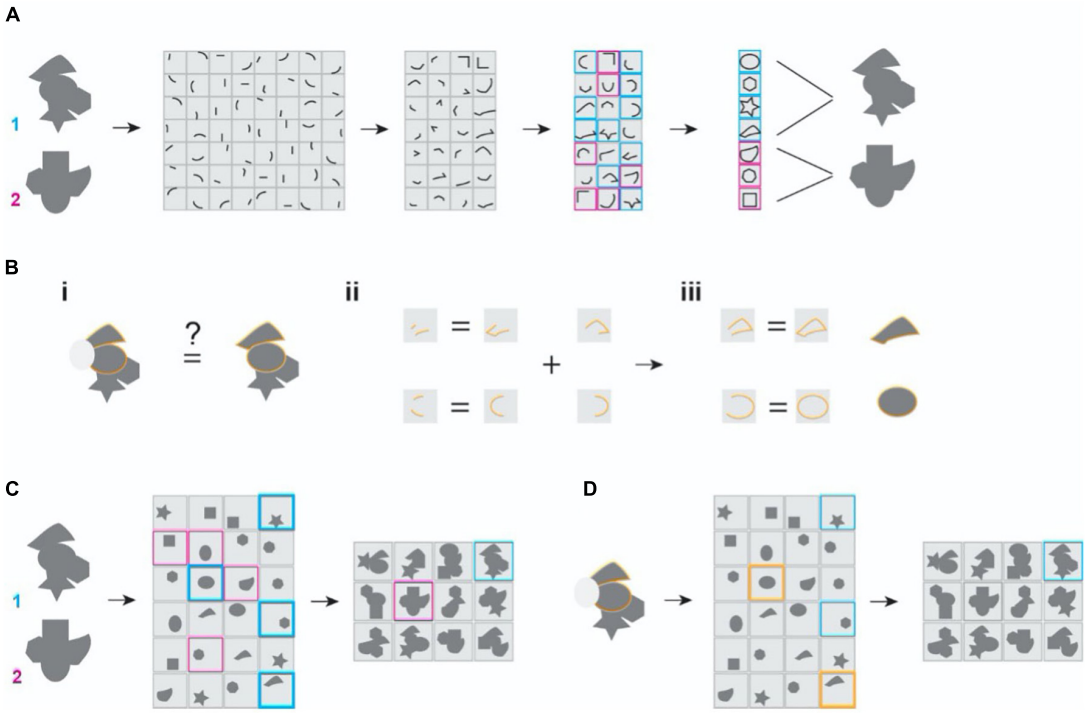

FIGURE 1 | Illustration of the classic and the maximal dependence capturing frameworks for object representation. **(A)** Classic framework. Inputs are decomposed into independent features and are reassembled hierarchically into more complex combinations to represent separate objects. Two randomly shaped objects are depicted. Their representations at later stages are color-coded. Gray patches depict the receptive field properties of representation neurons. **(B)** The problem of missing parts. **(i)** Occlusion (depicted as the gray oval obscuring the object) needs to be associated with the un-occluded one to be identified. **(ii,iii)** The classic models require learning the occluded feature or feature combinations to be the same at every stage of processing. There is also a need to learn every corrupted form to reach robustness. **(C)** In the MDC framework, coding units capture structural relationships among the features and encode them as a whole. **(D)** In the MDC model, missing feature (grayed out area) does not affect the encoding because redundancy allows the same units to represent the features and the objects even when parts are missing. There is no need to learn from all corrupted forms.

the two fields because both the brain and AI models adopt the same assumptions and framing.

One Problem, Disparate Solutions

To individual organisms, object recognition is for the purpose of determining the presence of an object. In computation models, the task has been divided into multiple problems for the brain to solve. The solution that addresses a specific problem often does not apply to others. For example, dividing the cognitive task into two distinct operations creates incompatible solutions. In the first operation, sensory features belonging to an object are hierarchically represented. Various features associated with the same object must be segregated from the background elements, and they need to be bound together (Von der Malsburg, 1995) (Riesenhuber and Poggio, 1999a; Singer, 1999). As such, the representation of objects also must solve the "binding" problem, which differs from the segmentation problem that assigns features to the proper objects when ambiguity arises (Treisman, 1999). Studies have suggested utilizing the temporal synchrony of neurons (Singer, 1999; Von der Malsburg, 1995) or non-linear maximum pooling of their activities to solve these segmentation and binding problems (Riesenhuber and Poggio, 1999a,b).

In the representation stages, perspective invariant representation is also to be achieved. The main model maps various perspectives of a 3D object to a stored standard view to achieve perspective invariance (Ullman and Basri, 1991; Ullman, 1996, 1998). However, at this stage, activities of view-tuned neurons are combined linearly as weighted summation (Poggio and Edelman, 1990; Poggio and Girosi, 1990) rather than non-linear maximum pooling proposed for the binding problem.

In the second operation, the process of discrimination and classification requires yet another set of rules. These rules are mostly associated with statistical learning, for example, manifold learning to disentangle representations of the same objects from others (Chen et al., 2018; DiCarlo and Cox, 2007). Moreover, the learning process requires labels for the classes. In the artificial networks, only the readouts at the final stages contain information to unambiguously classify or identify the objects (Krizhevsky et al., 2012; Goodfellow et al., 2016). This design does not have a parallel in the brain.

The framing of object recognition as a two-step process is problematic not simply because of the disparate solutions required. It has fundamental flaws in the assumptions. First, the various forms of input corresponding to the same object will have as many representations. There is no *a priori* label to tell

that these patterns belong to the same object and allow the type of learning in current neural network models. Second, even if such a mechanism exists, it requires storing class information of individual objects separate from the representations themselves, which is a problem in and of itself. Finally, it is unknown where in the brain the separation between representation and classification takes place.

The Conundrum of Hierarchical Assembly

Since Hubel and Wiesel first proposed hierarchical organization to explain the complexity of the receptive field properties observed in the visual pathway, the concept has become a cornerstone in understanding visual processing (Hubel and Wiesel, 1962; Marr, 1969; Marr and Nishihara, 1978; DiCarlo et al., 2012). Furthermore, studies of the visual processing streams have revealed shape-specific cells in high-order centers like V4 and IT. These cells are selectively tuned to faces or objects (Gross et al., 1969; Tanaka et al., 1991; Gross, 1992; Tanaka, 1992) and are involved in their recognition (Damasio et al., 1990; Damasio and Damasio, 1993). However, while physiological evidence is consistent with the hierarchical model, there is little anatomical evidence to demonstrate the progressive integration of elemental features along the hierarchy. Although cells in V1 have larger receptive fields than the retina, there is no obvious difference between V1 and V2 (Van den Bergh et al., 2010). In rodents, the receptive field is already large in V1, where the neurons' spatial tuning can be as large as 34 degrees (Van den Bergh et al., 2010). Nor is there strong evidence indicating that the cortical neurons perform stepwise integration. In fact, Felleman and Van Essen have argued that "there is no *a priori* reason to restrict the notion of hierarchical processing to a strictly serial sequence" and "any scheme in which there are well-defined levels of processing can be considered hierarchical" (Felleman and Vanessen, 1991). Indeed, it appears that each stage is reorganizing the input patterns for specific purposes (Bruce et al., 1981; Desimone et al., 1984; Tsao and Livingstone, 2008).

Nevertheless, many hierarchical models of object recognition rely on serial integration to achieve object representation (Fukushima and Miyaek, 1982; Anderson and Vanessen, 1987; Olshausen et al., 1993). These models share a conundrum with regard to how specific a cell should be in its response to sensory features. Experimental observations indicate that high-order neurons can be highly specific (Quiroga et al., 2005; Ponce et al., 2019). Many cortical neurons respond robustly to specific stimulus patterns, but slight changes in input could greatly reduce their responses (Rolls, 1984; Tanaka et al., 1990; Tanaka, 1993, 1996). If neurons are highly selective, the number of neurons needed to accommodate the possible feature combinations is astronomical. The improbability of the Grandmother Cells best illustrates this issue (McCulloch et al., 1959; Konorski, 1967; Barlow, 1995; Gross, 2002; Bowers, 2009). The concept of a grandmother cell refers to a neuron sitting at the top of a hierarchy to represent specific objects uniquely, even though it was initially raised as a singular addressable memory unit (Gross, 2002). While it is possible to create this type of highly selective cells, the requirement of generating specific cells that lead to the buildup of the grandmother cells is improbable.

In an alternate scenario, neurons can be less selective to avoid combinatorial explosion. Rather than relying on the highly specific responses, a population of less specific neurons can collectively encode the objects (Young and Yamane, 1992; Pasupathy and Connor, 2002; Chang and Tsao, 2017). However, it will be difficult to resolve ambiguities and distinguish similar input patterns in this arrangement. Recently it has been proposed that neurons at higher levels of visual processing do not represent specific feature combinations but encode individual axes in a high-dimensional linear space where each location corresponds to an object (Chang and Tsao, 2017). While such coding is possible, it may not be very effective in dealing with external and internal noises in the system. Deciding on the dimensionality of the object space creates another challenge. Moreover, there does not appear to be a need to code the entire space when only a few points in the space are relevant. This problem currently does not have a solution in the hierarchy model.

Problem With the Efficient Code

Most sensory neurons are ambiguous in representing the physical or chemical properties of stimuli. Photoreceptors and cochlear hair cells respond to a range of light and sound spectra, respectively (Russell and Sellick, 1978, 1983; Crawford and Fettiplace, 1980, 1981; Goldsmith, 1990; Rodieck and Rodieck, 1998). In the olfactory system, multiple odorants activate individual sensory neurons (Ressler et al., 1993; Vassar et al., 1993; Mombaerts et al., 1996; Treloar et al., 2002; Mombaerts, 2006; Fantana et al., 2008; Ma et al., 2012). In addition, the neurons are noisy, and their response changes as they adapt to stimulus intensity, duration, or context (Stockman et al., 2006; Rieke and Rudd, 2009; Rudd et al., 2009). Such response characteristics create confound in deciphering the precise stimulus. Early processing stages appear to mitigate this confound. One theoretical foundation of the early sensory transformations is efficient coding (Attneave, 1954; Barlow, 1961). Adapted from Information Theory, efficient coding has focused on minimizing redundancy in information transmission by encoding independent features present in natural stimuli (Laughlin, 1981; Olshausen and Field, 1996; Bell and Sejnowski, 1997; Lewicki, 2002; Smith and Lewicki, 2006). Individual neurons tuned to these features serve as independent encoders that efficiently relay information about the surroundings (Field, 1994; Olshausen and Field, 1996, 1997; Bell and Sejnowski, 1997). Models based on this theory successfully explain the receptive fields of neurons in the retina and the primary sensory cortices (Laughlin, 1981; Olshausen and Field, 1996; Bell and Sejnowski, 1997; Lewicki, 2002; Smith and Lewicki, 2006). Also, as the theory predicts, the response properties of neurons in the retina, the thalamus, and the primary visual and auditory cortices conform to the statistics of natural stimuli (Barlow et al., 1957; Hartline and Ratliff, 1972; Srinivasan et al., 1982; Atick and Redlich, 1990; Dong and Atick, 1995; Olshausen and Field, 1996, 1997; Bell and Sejnowski, 1997; Simoncelli and Olshausen, 2001; Geisler, 2008).

However, these results are not without controversy. Neuronal recordings have revealed overlapping receptive fields and synchronized activity among the retinal ganglion and V1 cells in many species (Meister and Berry, 1999; Nirenberg et al., 2001; Reich et al., 2001; Puchalla et al., 2005; Pillow et al., 2008; Ohiorhenuan et al., 2010). Recent rodent studies show large spatial tunings of V1 cells and a much less organized primary visual cortex (Martinez et al., 2005; Niell and Stryker, 2008). In the mouse olfactory cortex, any given odorant activates multiple neurons not specific to the odorant (Poo and Isaacson, 2009; Stettler and Axel, 2009). These observations do not conform to efficient coding. They suggest a high correlation among neurons and redundant information transmission. Indeed, the presence of redundancy in neuronal responses has led Barlow to suggest that redundancy is useful for encoding object identities, although he has not proposed how the information is used (Barlow, 2001).

The efficient coding hypothesis, in fact, poses serious problems for cognitive robustness. An object is not merely a collection of features. The relationship among its features defines it. Encoding independent features remove information about these relationships from the sensory input. Consequently, the system faces a challenge to recover and store this information. The task becomes exceedingly difficult when occlusion, internal or external noise, or inactivity of neurons causes ambiguity in the input signal. Any inference of the absent fraction of the signal or the input is impossible without the relational information. Mechanisms like pattern completion can help in certain situations (Rao and Ballard, 1999; Lee and Mumford, 2003). However, these mechanisms require storing the relational information, a problem that does not have a ready answer.

Unsolved Problems With Statistical Learning

To achieve robustness, the current frameworks of object recognition rely heavily on statistical learning or post-representational inference (**Figure 1B**). For example, a 3-D object can be aligned and associated with its multiple 2D views (Biederman, 1987; Ullman and Basri, 1991; Ullman, 1996). Similarly, corrupted and occluded inputs can be linked with their non-corrupted forms along a manifold through learning (DiCarlo and Cox, 2007). As such, recent neural network-based models utilize extensive training using many examples to identify objects from incomplete images or novel perspectives. Presumably, such comprehensive training reveals features and their relationships that facilitate robust recognition (**Figures 1Bii,iii**).

However, the robustness provided by statistical learning is *retrospective*, meaning that only the learned examples, or those closely resembling them can be identified with high accuracy. On the other hand, the animal brain can recognize objects with *prospective robustness*, which we define as the accuracy in identifying novel objects and their different forms without additional learning. Animal brains learn a new object and recognize it without having to experience all of its variations in form. This ability to use few examples and perform prospective recognition is missing in models based on statistical learning.

NEW PERSPECTIVES OF OBJECT RECOGNITION

Given the caveats of the current framework, we wish to offer new perspectives of object recognition. Specifically, we intend to establish a framework that considers the animal's ethological needs and affords both retrospective and prospective robustness.

Before discussing the details of the framework, a short note on "objects" is necessary. Physical "objects" that we see and recognize are explicit in visual inputs. However, other sensory modalities also signal "objects" that serve similar functions. For example, an acoustic object comprises a specific combination of sounds with characteristic frequencies and durations (Griffiths and Warren, 2004). A distinct blend of odorants constitutes an odor object (Keller et al., 2007; Barwich, 2014, 2018, 2019; Smith, 2015). These "objects" are amorphous, but they signal the presence of their emitters. Further, an animal can trace them through directionality or concentration gradients. We believe that any discussion on object recognition must include these objects and should not just concern the visual ones.

The Ethological Perspective

To understand object recognition, it is imperative to consider its behavioral and evolutionary purpose rather than the accuracy in representing the physicochemical properties of the stimuli (Burge, 2010). An object is meaningful to an animal when it is informative of its associated consequences. Accordingly, recognizing the presence of an object through its *identity* is paramount to the behavioral consequences. Once the animal establishes the object's *identity*, it can act appropriately to increase the chances of its survival. Thus, the core objective of recognition is to determine objects' identities to trigger proper behavior.

How does one determine object identities? All object features do not convey the identity information equally. Some features or feature combinations are more useful than others in identifying objects. Sensory organs capture information redundantly, which can be useful in eliminating input ambiguity in object identities. Accurately representing every physiochemical property of the stimulus may not be necessary. From this perspective, an animal does not need to know every detail about the object to identify it correctly.

In his seminal study, Barlow has described the "on-off" retinal ganglion cells in the frogs as "the detectors of snap-worthy objects" (Barlow, 1953, 1961). He argued that a prey fly at a reachable distance from the frog would exactly fill the receptive field of these cells and generate the most vigorous response in them (Barlow, 1953). Thus, the frog visual system can detect flies without representing every detail and reliably set off the hunting sequence. Indeed, a later study further characterizing these cells' properties revealed hallmarks of perception rather than simple sensation (Lettvin et al., 1959).

However, the current models have lost this initial insight. While the idea of representing object identities without the

specifics is not entirely new (Marr and Nishihara, 1978; Barlow, 1989; Logothetis et al., 1994; Marr, 2010), the principal focus in current models is on accurate feature representation and reproducing the receptive field properties of the cells. Representation of object identities is pushed to the top of the hierarchal representation structure, and the hierarchy and deep structures have become obligatory for object recognition.

Arguably, object representations at different levels of hierarchy serve different purposes. If a part of the nervous system can accumulate sufficient information from the features or combination of features to guide behaviors, it has achieved object recognition. It does not need representation at the top of the hierarchy. Frogs can snag prey using just "on-off" cells in their retina. Moreover, there is no need for processing through a deep structure of neural networks. Relatively shallow structures process odor information in species across phyla. The insect and mammalian olfactory systems have two processing stages: the antennal lobe and the mushroom body in insects, and the olfactory bulb and the olfactory cortices in mammals. This shallow structure nonetheless allows robust recognition of odor objects to guide behaviors (Kobayakawa et al., 2007). Lastly, the early stages are not required for processing at the later stages. For example, the primary auditory cortex is not required for speech recognition (Hamilton et al., 2021). Thus, encoding object *identity* does not require an accurate representation of stimulus *features* along a hierarchy.

An Information-Theoretical Perspective

What is sufficient information to determine object identity? Multiple objects often share the same features; therefore, these features cannot uniquely identify an object. The uniqueness of an object resides in the specific combination and the structural relationships among these features (Hoffman and Richards, 1984; Biederman, 1987; Hummel and Biederman, 1992). For example, recognizing a predator in various concealments or camouflages is possible because the relative configuration of the colors, spots, and shapes, though only visible partially due to the concealment, is adequate for its identification. In other words, the information sufficient for recognizing an object is embedded in feature combinations that considerably reduce the uncertainty about it.

Notably, for a given object, there can be multiple feature combinations that uniquely identify it. From the information-theoretical perspective, these feature sets provide redundant information about the object, which can be useful in resolving ambiguities. Therefore, any sensory processing framework must address encoding the most informative structural components and using redundancy to generate consistent object representations across different experiences.

To better understand the informativeness of features about objects, suppose we need to identify N objects, each defined by a unique combination of M features. For a given object O_i, let its full set of features be f_{full}. Assuming a condition where only a subset f_{subset} of f_{full} is available to the system, the following relation with regard to the entropy H holds:

$$H\left(O_i \mid f_{subset}\right) = H\left(f_{full} | f_{subset}\right) = H(f_{full}) \; - \; I\left(f_{full}; f_{subset}\right) \quad (1)$$

i.e., the uncertainty in predicting the full set of features or the object given an input subset decreases as the mutual information (I) between the given subset and the full set increases. This mutual information is the informativeness of the given subset of features about the object. From this information-theoretical perspective, *object recognition is achieved when the subset of features is maximally informative and uncertainty about the object vanishes*. At this point, the subset unambiguously informs the brain of the object's presence. Multiple feature combinations can be equally informative about the object; therefore, multiple ways of representing the same object based on these combinations are possible.

With this definition, we propose that the problem of object recognition be stated as finding the computational rule that allows the neurons to robustly encode object identities using the most informative feature combination. We next develop this idea to propose the maximal dependence capturing principle and provide a mathematical solution to the problem.

The Maximal Dependence Capturing Principle

We can express the informativeness of features as the information provided by the representation units (the encoders). If there are K encoder $\{x_1 \ldots, x_j \ldots, x_K\}$, each capturing a different substructure of the object O, then the mutual information between encoder j and the object $I\left(O; x_j\right)$ is the same as the informativeness of the sub-structure about the object. In noiseless conditions

$$\sum_j I\left(O; x_j\right) \; \geq \; I\left(O; X\right) \quad (2)$$

where X is the representation of the object.

This relation shows that in a noiseless situation the sum of all information about the object from individual encoders will be more than the information about the object from its entire representation. This extra information can help resolve ambiguities. Further, fewer encoders can be sufficiently informative when the system maximizes the mutual information between the object and individual encoders. The higher the mutual information, the fewer the encoders are needed, and the more robust the encoders can be in identifying an object.

Most frameworks of sensory processing consider the encoders to be independent. This arrangement minimizes the mutual information between them, creating situations where individual encoders are not informative of any object. In contrast, we suggest that individual neurons capture maximum information about individual objects and are not independent. We refer to this as the maximal dependence capturing (MDC) principle because mutual information measures dependence. Further, we propose sensory processing to follow this principle.

This principle, as we show, leads to a sensory coding framework that can enable robust object recognition (**Figures 1C,D**). In this framework, neurons do not serve as independent encoders but encode the structural relationship among sensory features. It is not difficult to see that if a neuron captures the entire object structure, it can uniquely represent

it. Other neurons may detect the object's substructures and convey redundant information, but they are not needed. Using a mathematical model, we demonstrate that a specific form of sparse coding enables capturing such dependence and generates unique representations of the same object in various forms. This framework can achieve prospective robustness without the requirement of deep layers or statistical learning. We show that the framework is adaptive to object sets and can naturally lead to simple cell-like receptive field properties that have been characterized as independent encoders.

From this framing, several disparate problems can be treated as the same. Corruption and occlusion, for example, are problems of recovering full identities from partial signals. If neurons encode the dependence between the visible features and the object, missing parts can be inferred from the dependence to effectively solve the occlusion problem. As the dependence does not change with scale or location, the corresponding invariances can be achieved. The same holds even for 3-D object representations. Since each 2-D view can be considered a unique combination of a subset of an object's features, multiple 2-D views can provide the same identification. Therefore, the view angle problem becomes whether a particular view of an object contains sufficient information to determine its identity. From the information-theoretical perspective, the solution to these disparate problems is the same: capturing the most information about the object as feature combinations. Moreover, as we show below, the activity of the most informative units can be maintained the same even when the input pattern is incomplete. This characteristic removes the inference requirement, thereby allowing the same set of rules to accommodate both representation and classification.

THE MODEL

We find that during the linear transformation of input patterns to representations, individual neurons can get tuned to comprehensive input structures if we constrain the representations to be sparse and make the transformation process non-negative. The sparsity constraint ensures that any object representation comprises a minimal number of active neurons and is maximally distinct from others. Non-negativity in transformation prevents encoding inputs as differences of structures. This type of coding eliminates the chances of neurons getting tuned to superpositions of multiple objects. Also, a non-negative representation is biologically more meaningful because action potentials are positive signals.

The specific objective function that we optimize to attain the desired transformation for a finite set of objects is

$$\underset{\Phi,\mathbf{A}}{argmin}\ ||\mathbf{X}-\Phi\mathbf{A}||_2^2 + \lambda||\mathbf{A}||_1 \ subject\ to\ \Phi \geq 0\ and\ \mathbf{A} \geq 0 \quad (3)$$

Here X is a matrix of inputs, and A is a matrix of corresponding representations. Φ is a matrix with columns corresponding to the tuning properties of the representational neurons. It serves as the basis set for representations and comprises the inputs' informative structures. We refer to it as the *dictionary* according to convention.

The first term in the objective function measures the difference between the input structure and the structures captured in their representations. Minimizing it ensures that the representations reflect most of the input structures. The second term is a measure of representation sparseness. It serves as a penalty on the total activity of neurons in a representation. Note that the intention is to reduce the number of active units in a representation, i.e., the l_0 norm, to minimize representation overlap. Mathematically, however, an analytical solution to l_0 minimization is not possible. Therefore, we minimize neurons' total activity in representations (the l_1 norm of representations). Also, note that we set the representation dimension to be larger than the inputs to achieve sparseness. In this setting, the transformation conforms to the observation that higher brain centers often possess several folds more neurons than the sensory organs.

The optimization function is similar to those employed to capture the natural scenes' independent components (Olshausen and Field, 1996, 1997; Simoncelli and Olshausen, 2001; Geisler, 2008). However, we use this objective function in a different context. Instead of identifying the independent features based on the statistics of the entire input space, the goal here is to represent a limited set of inputs by capturing their most informative structures. Furthermore, the non-negativity and the sparsity constraints force individual neurons to tune to comprehensive input structures. Without non-negativity, the tuning properties of neurons can be arbitrarily complex (Olshausen, 2013).

Dependence Capturing

We illustrate this framework's key characteristics by encoding binary symbol images from world languages (**Figure 2** and **Supplementary Figure 1A**). With 256 input and 800 output units, the representations in these simulations are dimensionally expanded (**Figure 2A**). Learning to represent 1,000 symbols results in dictionary elements containing local and global structures (**Figure 2B**). Structures of individual dictionary elements contain varying information about different inputs. The localized structures are less informative about any input, but the comprehensive structures are unique to specific inputs and contain the most information about them (**Figure 2C**). We plotted the histogram of the dictionary elements' maximum information contents for any input (**Figure 2D**). The histogram is heavily skewed toward larger values, indicating that the framework successfully captures the highly informative structures. An interesting observation in these simulations is that multiple dictionary elements are structurally similar. For example, several dictionary elements shown in **Figure 2B** have the same oval-like shapes. These dictionary elements are utilized in distinctively representing very similar input symbols (**Figure 3A**). These symbols have subtle differences, which are captured in these matching dictionary elements. Thus, the finding demonstrates that the MDC framework does not just capture the distinguishing structures. It allows the redundant encoding of features shared by multiple objects. The computation in the framework can successfully extract complex structures

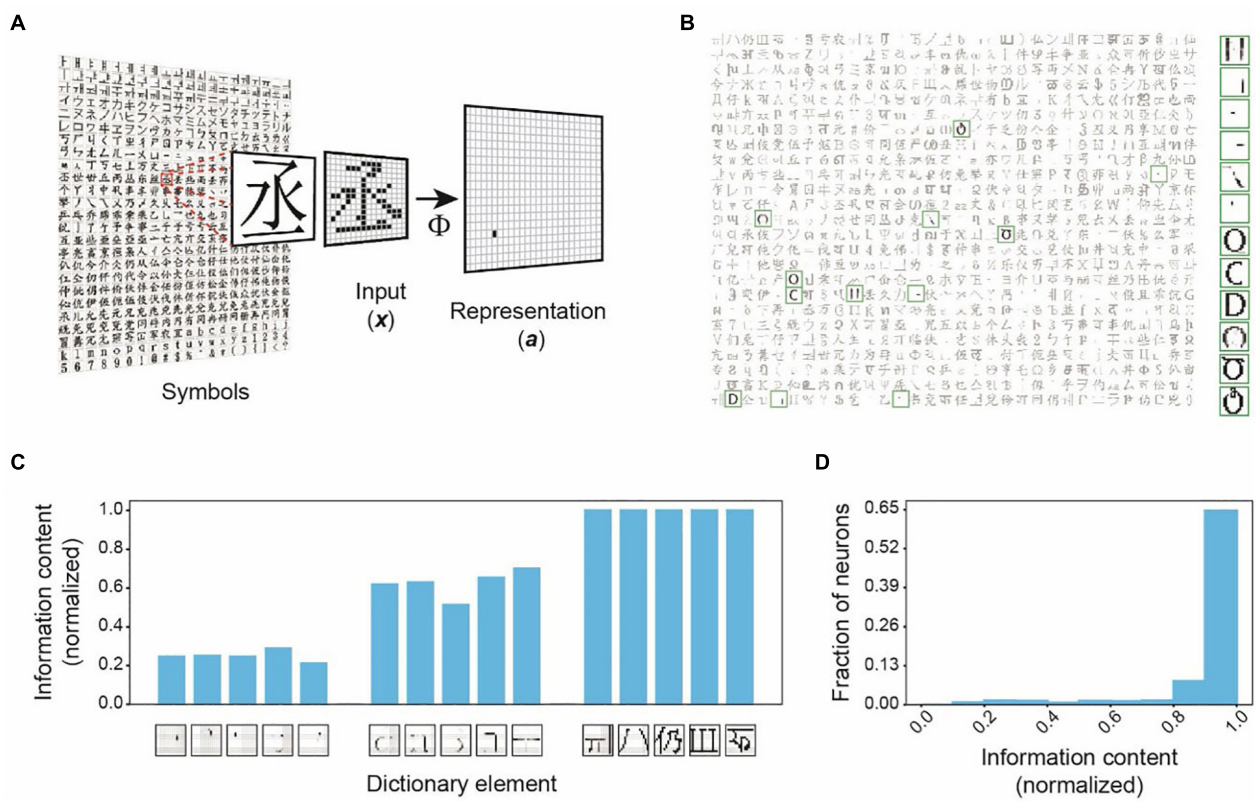

FIGURE 2 | Dependence Capturing by the MDC framework. **(A)** Illustration of encoding in the MDC framework. Symbols from world languages are converted to 256 (16 × 16) pixel images (*x*) that are transformed into the activity of a set of 800 representation units (*a*) to encode symbol identities. An example of the process is shown using a character encoded by a single representation unit to indicate the ability to encode complex structural features. **(B)** Structures of the dictionary elements (receptive fields) learned from 1,000 symbols. Highlighted elements displayed in the larger size are most active while representing the inputs shown in **Figure 3**. Note the similarity among some of the elements. **(C)** Information contents of sample dictionary elements normalized to the maximum observed information content. While very simple structures are least informative about any object, more comprehensive structures are highly informative. **(D)** Distribution of the normalized information contents of the dictionary elements. Most of the structures are highly informative about specific objects indicating that the mathematical framework captures features that share maximum dependence with the inputs.

naturally present in the stimuli without creating arbitrary or overly complicated dictionary elements.

Prospective Robustness in Invariant Representation

We next test whether the MDC framework can encode objects distinctively, especially in conditions of noise and corruption. In classic frameworks, neural network models require deep layers to enhance robustness. The MDC framework captures comprehensive input structures, and we expect it to be robust against corruption without the deep layers. To obtain the representations of corrupted input patterns using the learned dictionary, we optimized the following objective function under the non-negativity constraint:

$$\text{minimize}||\mathbf{a}||_1 \text{ subject to } ||\mathbf{x}-\Phi\mathbf{a}||_2 \leq \varepsilon \qquad (4)$$

The optimization ensures that the MDC framework utilizes the same computational rule for learning to obtain representation. Thus, it distinguishes itself from the hierarchical assembly, where the learning rule is separate from the transformation. This approach also contrasts with the previous approaches that use direct convolution of input with the dictionary to generate representations (Olshausen and Field, 1997; Rao and Ballard, 1999; Rozell et al., 2008; Lörincz et al., 2012).

The two-layer model based on our framework readily distinguishes highly similar patterns and represents them differently (**Figure 3A**). Representation neurons have minimum correlation (**Figure 3B**), and the symbol representations are sparse (**Figure 3C**). Whereas the correlation matrix of representation units is very close to identity, there are similar non-zero correlations among pixels in both the input and the dictionary (**Supplementary Figures 1B,C**). Resembling correlations indicate that the dictionary captures complete input structures.

Moreover, this framework achieves the desired invariance in representation. Without learning from corrupt examples, the model can transform inputs corrupted by Gaussian noise (**Figure 3Cii**), randomly missing pixels (**Figure 3Ciii**), or partial

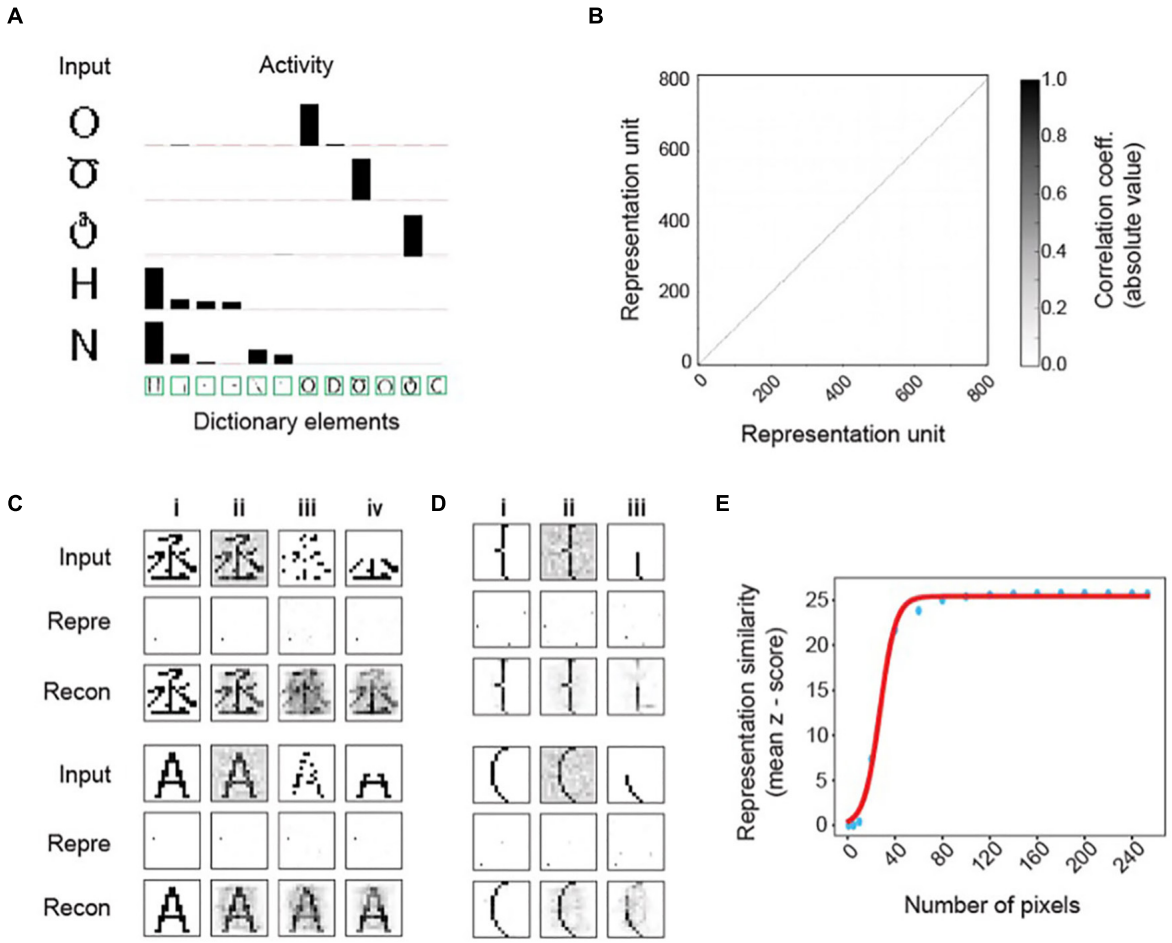

FIGURE 3 | Invariant object representations in the MDC framework. **(A)** Representation of inputs by the activity of the dictionary elements. Only the most active ones are shown. The height of the bars indicates activity evoked by the input patterns. Note that the inputs only activate the dictionary elements that are most similar to them despite the similarity among the elements. Such coding results in very sparse input representations. **(B)** Correlation among representation units is minimal, and the pairwise correlation matrix of the representation units is an identity. **(C)** Representations of original **(i)** and corrupted inputs under noisy **(ii)**, pixel loss **(iii)**, and occlusion **(iv)** conditions. Representations of the corrupted signals (Repre) are similar to those of the originals. Reconstructed images (Recon) from the output units resemble the original symbols. **(D)** Example of two highly similar symbols being distinctly and robustly represented. The original input signals **(i)** corrupted by noise **(ii)** or occlusion **(iii)** are transformed to output activities that are similar to each other. Reconstructed images recover the original signals. **(E)** Z-scored similarity (specificity) between representations of corrupted and original signals as a function of total pixels in the input layer [randomly selected as in panel **(Ciii)**]. Scores calculated with different pixel numbers (blue dots) is fit to a sigmoidal curve (red line).

occlusion (**Figure 3Civ**) to representations identical to those of the uncorrupted input forms (**Figure 3Ci**). Reconstructing images by linearly combining dictionary elements restores whole symbols rather than parts (**Figure 3C**). Using the Z-score of the pairwise cosine distances between the representation of corrupted inputs and all learned symbols, we observe high specificity for the correct input-symbol pairs, indicating that the framework generates precise representations. Notably, the representations are sensitive to small differences in the input patterns. For example, two highly similar input patterns are represented differently and robustly under various corruptive conditions (**Figure 3D**). Monte Carlo simulations with randomly missing input units yield highly specific representations with as few as 60 (23.4% of the 256) units (**Figure 3E**). Thus, the computational framework fulfills the requirements of stability and sensitivity set forth by Marr and Nishihara (1978) without requiring deep layers or learning from many variable examples. Importantly, this example shows prospective robustness, as the stable representation of symbol identity does not depend on statistical learning through many corrupt examples.

A Robust Code for Face Recognition

We next tested the two-layers model in encoding complex, non-binary signals such as human faces (**Figure 4**). We trained the model on 2,000 human faces (**Figure 4A**). The learned dictionary elements are composed of a complex assemblage of facial features, again suggesting that the algorithm captures complex structures present in the training set (**Figure 4B**). The face representations are stable, unique, and robust against common alterations such as headwear, facial hair, or eyewear (**Figure 4C**). The model

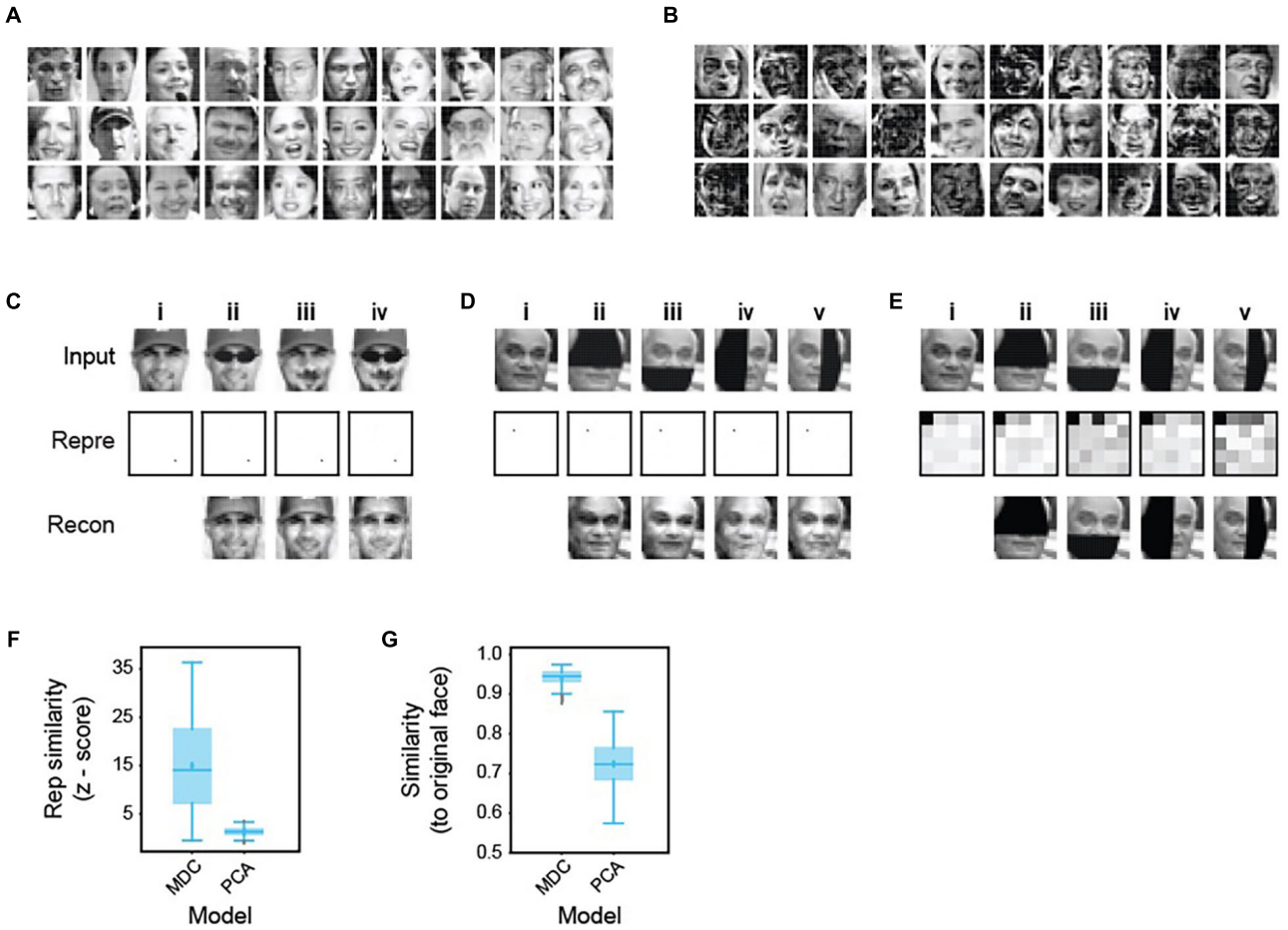

FIGURE 4 | Robust representation of human faces. **(A)** Examples of face images in a library of 2,000 (1,000 male and 1,000 female) used to train a two-layer model. **(B)** Examples of dictionary elements learned from the face library. Note that they incorporate complex combinations of facial features but are not necessarily part or whole of any specific face. **(C)** Representation and recovery of faces with different alterations. A face **(i)** was altered to wear a pair of sunglasses **(ii)**, a mustache **(iii)**, or both **(iv)**. The altered faces' representations were nearly identical to the original, even though these examples were not in the training set. Images reconstructed from the representations based on the dictionary were similar to the original images. **(D)** Representation and recovery of occluded faces. Four different occlusions of a face—top **(ii)**, bottom **(iii)**, left **(iv)**, and right **(v)** were generated. Representations of the occluded faces were highly similar to the original one **(i)**. Reconstructions also matched the original face. **(E)** Face identity was not preserved in representations based on PCA performed on the 2,000 training faces. Representations of original **(i)** and occluded faces **(ii–v)** were obtained in the principal space (the first 25 components are shown to highlight the differences). Representations of occluded faces are different from the original. Reconstructed images match the corrupted rather than original faces. **(F,G)** Specificity (Z-score similarity) of face representations **(F)** and cosine similarity between reconstructed and original images **(G)** were calculated for MDC and PCA models. 50 faces were chosen randomly from the training set to create the four occluded versions.

produces nearly identical representations while transforming the same face with a mustache, a pair of sunglasses, or both. It maintains representation consistency even when we block half the input face in different positions (**Figure 4D**). Inversely reconstructed images from the representations are similar to the unadulterated faces (**Figures 4C,D**). Notably, the model achieves robustness from learning only 2,000 examples and without using any corrupted images. Achieving such consistency is in direct contrast to many approaches using variegated examples as training sets.

Moreover, the dictionary learned from the training set can be applied to an entirely new set of faces. We used it to obtain representations of facial images in the Yale face base, which contains 15 individual faces, 11 different lighting conditions, and facial expressions. The representation correctly categorized the faces according to the individuals (**Supplementary Figure 2**).

We compared our code against a recently proposed principal components-based face code (Chang and Tsao, 2017). Using dictionaries obtained through principal component analysis (PCA), the same face with different parts occluded generated different representations. Image recovery produces occluded but not uncorrupted images (**Figure 4E**). In contrast, images recovered from the MDC representations of corrupted inputs are similar to the original ones (**Figure 4C,D**). Quantification of specificity using Z-scores shows that our model generates highly specific representations (**Figure 4F**).

PCA-based decoding does not exhibit such selectivity or similarity (**Figures 4F,G**). Thus, a face code based on the MDC framework is robust against corruption, whereas the one proposed before is not (Chang and Tsao, 2017; Stevens, 2018).

Adaptive Nature of the Maximal Dependence Capturing Code

In the MDC framework, an increase in the number of objects can influence the dictionary elements' structures. We explored the relationship by varying the numbers of encoded inputs (N) and representation neurons (K). With N fixed, at low K, dictionary elements are more localized (**Figure 5A**). The representations are less sparse and more active neurons encode the same symbol (**Figure 5B**). Increasing the number of neurons makes the response sparser.

Conversely, with a fixed K, an increase in the number of inputs causes a divergence of the dictionary element structures from the structure of inputs (measured using the K-L divergence (Kullback and Leibler, 1951) of pixel distribution between the input and the dictionary; **Figure 5C**). As more representation units encode each symbol, redundancy among them also increases (**Figure 5D**). At high N, the redundancy approaches that of the input, suggesting a decreased efficiency. The same observation holds for complex signals. The dictionary elements for faces are more localized at low K values, resembling local facial features (**Figure 5E**). At high dimensions, they become more complex and face-like (**Figure 5F**). More units are required to encode each face at lower dimensions (**Figure 5G**).

The MDC framework's premise is that individual encoders maximally capture structures from the input signals, which results in complex dictionary elements. This characteristic appears to be contradictory to the localized receptive fields observed in primary sensory cortices. While the efficient coding hypothesis explains simple tuning structures as the independent components of natural inputs, we tested whether the MDC framework could also produce these localized receptive fields.

In our simulations, the tuning properties of neurons shift from being comprehensive to localized as the network encodes more objects. This trend suggests that the MDC framework captures simpler features when forced to encode more objects. In the visual system, the primary visual cortex relays all visual information. So, it must accommodate all visual objects. We reason that the overwhelming number of visual objects can force the cortical cells to adapt to the natural statistics of the visual input and tune them to the localized features. To test our reasoning, we trained the network with varying numbers of image patches taken from natural images (Van Hateren and van der Schaaf, 1998) (**Figure 6A**). Our method is like the approach of Olshausen and Field (Olshausen and Field, 1996; Olshausen and Field, 1997) but with non-negative constraints and a more limited number of training inputs. We parsed the images into positive and negative signals to simulate the On and Off channels in the mammalian visual system.

Training with up to 30,000 patches develops local, simple cell-like, and complex dictionary elements (**Figure 6B**). With a low number of training images, the dictionary elements are relatively complex (**Figure 6C**). Despite high correlations among the images, we observe minimal correlation among representation units (**Supplementary Figures 3A,B**). Increasing the training set size produces more dictionary elements with localized and orientation-selective projective fields similar to the receptive fields of simple cells in the mammalian primary visual cortex (Hubel and Wiesel, 1962) (**Figures 6B,C**). The simple cell-like dictionary elements resemble Gabor filters, as found in earlier studies (Field, 1987; Jones and Palmer, 1987).

Interestingly, complex receptive fields persist in all training conditions. The percentage of simple projective fields in the dictionary increases with the size of the training set. With a fixed set of training images, an increase in the encoding dimension reduces correlation among the encoding units but increases the correlations among dictionary elements (**Supplementary Figure 3C**). Thus, simple receptive fields conforming to the classic interpretation emerge when the model encodes many natural images under the MDC framework (Laughlin, 1981; Atick and Redlich, 1990; Dan et al., 1996; Lewicki, 2002; Smith and Lewicki, 2006). Importantly, complex tuning is always present in our model without any synthesis from the simple cells, as the classic model predicts (Hubel and Wiesel, 1962).

DISCUSSION

We propose that maximal dependence capturing is a general principle of sensory processing and an alternative to the redundancy reduction principle. The primary motivation behind redundancy reduction is to arrive at a factorial code for object representation to optimize information transmission (Barlow, 1961, 1989). However, a factorial code based on independent features is unsuitable for invariant coding, especially in corruption or occlusion cases. Hierarchical models learn complex feature combinations to achieve robustness and no longer use independent components for representations. With this "reduce and capture" strategy, where the model reduces redundancy among features before pooling them together by deep networks, the classic framework is self-conflicting and creates unnecessary problems. The MDC framework resolves the conflict with a "capture and reduce" strategy for redundancy. By assuming that dependence capturing is the essence of the sensory system, the coding units extract the most informative combination of features that uniquely identify specific objects. Though it makes representation redundant, sparse coding ensures that minimum correlation exists among the coding units. Thus, the framework obviates the need for a hierarchical assembly to associate features. It embeds individual features in complex dictionary element structures. Moreover, since the dictionary captures the dependence *a priori*, the framework does not require multiple corrupted forms to learn the associations.

The MDC framework's unique characteristic is that the receptive fields capture highly informative structures about individual objects. With this characteristic, tuning of individual units can be very similar and correspond to multiple objects. As a result, correlations may develop in their responses.

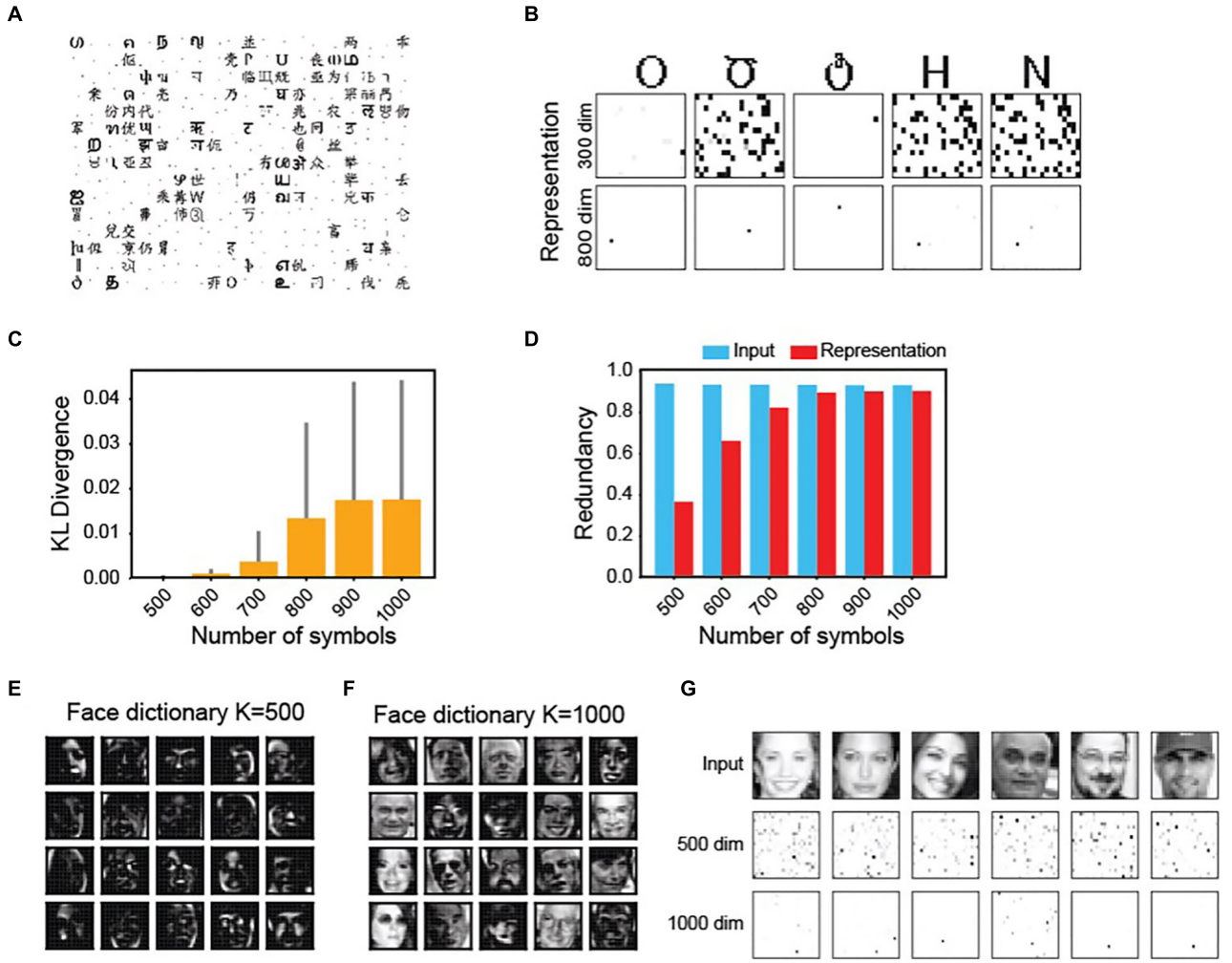

FIGURE 5 | Complexity of tuning properties is determined by the number of objects and dimension of the representation. **(A)** The structures of dictionary elements for the symbols in **Figure 2** with a 300-unit output layer. Compared with that of the 800-unit shown in **Figure 2B**, there are more localized features. **(B)** Representation sparsity increases with increased dimensions. **(C)** K-L divergence between pixel distributions in the input signal and in the dictionary element as a function of the number of symbols to be encoded. **(D)** Coding redundancies in input and output (representation) while encoding increasing numbers of symbols. **(E,F)** dictionary elements of faces with a 500-dimension **(E)** or 1,000-dimension **(F)** encoder-set. **(G)** Response elicited by faces are sparser with increased dimensions.

This characteristic appears to be antithetical to the notion of redundancy reduction, which demands individual encoding units to be as independent as possible. However, representations achieved under the MDC framework also satisfy a sparsity constraint. This constraint shrinks the activities of competing neurons and decorrelates individual units' responses even when their receptive fields have high levels of overlap. Thus, the MDC framework allows the efficient encoding, but the coding is for object identities. It is done by capturing complex features.

The MDC framework makes specific predictions that can be tested through anatomical and physiological experiments. For example, it predicts that the connectivity required to achieve the orientation specific tuning of early neurons should be less organized than previous models predict (Hubel and Wiesel, 1962, 1968; Fukushima and Miyaek, 1982; Anderson and Vanessen, 1987; Olshausen et al., 1993; Rolls and Milward, 2000). Decorrelation among neurons through lateral connections is an essential feature of the MDC framework. Shutting down inhibitory lateral connections is expected to reveal highly overlapped receptive fields among the neurons. On the contrary, models based on connectivity alone would predict a much smaller degree of receptive field expansion. Moreover, while inhibition-mediated decorrelation is a common feature in the nervous system, our model would predict that cells with similar tuning properties are likely to have stronger mutual inhibitory connectivity. Manipulating the lateral connections would give insight about this prediction.

We have shown that the robust representation of object identity does not require deep network structures. Thus, this framework can explain robust object recognition in animals with less complex brain or sensory systems that do not possess complex hierarchical organizations. Nevertheless, hierarchical

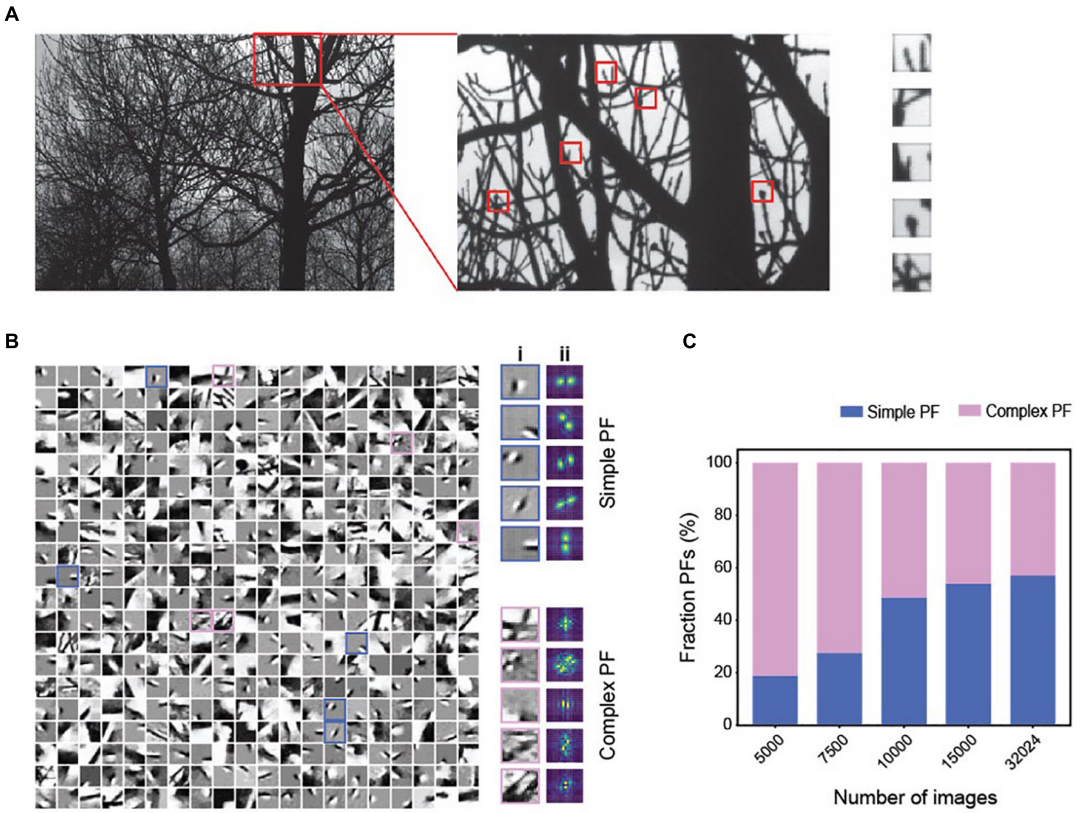

FIGURE 6 | The emergence of simple and complex receptive fields from training with natural images in MDC. **(A)** Illustration of image patches derived from the natural scene. **(B)** Examples of structures of individual dictionary elements. Both simple (cyan) and complex (magenta) elements are observed **(i)** and can be determined using Fourier transforms of the structures **(ii)**. **(C)** As the number of training images increases, encoders' tuning properties become more localized, and the percentage of simple PFs increases.

organization can arise to deal with increasingly complex stimuli during evolution. We suggest that as the animal needs to identify more objects, early processing can shift to encode localized features resembling the independent components. As we show, there is a relationship between the number of encoded objects and the complexity of the tuning properties, which allows both simple and complex receptive fields to develop under the same rule. When each neuron encodes the local association of features, multiple neurons are required to encode individual objects. The MDC framework allows subsequent levels to capture the global dependence among the local feature combinations. From the evolutionary perspective, the sensory systems have evolved to detect ethologically relevant signals. Analyzing environmental stimuli and parsing them into components of minimal redundancy is not necessary for this goal.

In a way, the MDC framework produces sparse distributed representations, which can account for some experimental observations, including the appearance of "grandmother cells" (Quiroga et al., 2005; Bowers, 2009; Rey et al., 2020). However, the previous models have mainly focused on reproducing the key features of neuronal tuning (Olshausen and Field, 1996, 1997; Bell and Sejnowski, 1997) and memory storage (Laurent, 1999; Palm, 2013). They do not provide a strong explanation for the change in the complexity of tuning properties along the processing stream. The MDC framework neither requires hierarchical assembly nor an account of all possible feature combinations. Since the encoders only capture the naturally occurring structure, the generation of "grandmother cell" like representation is a feature of the MDC framework. The cells, however, are not the traditional sense of "grandmother cells" because they do not sit at the top of any hierarchy.

A point worth noting is that the receptive fields of individual coding units resemble the objects themselves when the representations are sparsest. One may consider the coding scheme using these cells as a form of template matching (Burr, 1981; Buhmann et al., 1990; Yuille, 1991; Brunelli and Poggio, 1993). As we have shown, however, the MDC framework is not template-matching. The receptive fields may resemble whole structures of the objects, but they are not identical. Moreover, these receptive fields change as the coding units adapt to different numbers of inputs.

In this study, we have shown invariant representation for corrupted or occluded input patterns. Representation invariance takes many forms. Invariant representations result from scale, translational, and affine transformations are common. Although we have not explored these forms, we

believe the framework will allow easy incorporation of these transformations without evoking additional mechanisms. As the name of the MDC principle indicates, dependencies among the features are captured by neurons. The dependencies, and the informativeness of features, do not change with linear transformation. Thus, the framework will naturally generate covariant receptive field properties that are thought to enable invariant representation of the visual stimuli at higher levels of processing (Lindeberg, 2013, 2021).

Likewise, understanding temporal dependencies between objects and events and the ability to predict future is fundamental to survival of the organisms. More recent studies have focused on efficient representation of moving images (Rao and Ballard, 1997; Rao, 1999; Srivastava et al., 2015), whereas some have produced models encoding features that best predict future (Bialek et al., 2001; Palmer et al., 2015; Salisbury and Palmer, 2016). Although we have not explored the effectiveness of the MDC framework in capturing temporal redundancies, it is in principle achievable. At the minimum, the framework can be combined with other models to incorporate temporal dependence capturing. For example, Singer et al. introduces a two layered feedforward network to predict future frames of movies (Singer et al., 2018). It should be straightforward to achieve invariant representations of static frames, which can be utilized together with the Singer model to predict future frames more accurately. Alternatively, a unified model, which may involve multiple layers, can capture not only spatial but also temporal dependencies to predict environmental stimuli in the spatiotemporal domain.

In summary, this work offers a novel perspective of object representation. We propose that the sensory system should utilize the most informative structures from objects as the basis of their representations. The maximal dependence capturing principle allows neurons to capture these structures by learning the relationships among features that identify the objects. This type of learning eliminates the need for an analytical step to break down objects into its composing features and the need of the classic hierarchical assembly where brain representations of objects are built sequentially from their elementary features. Learning is possible without the deep structures or large training set. These characteristics of our framework make it generalizable to any system irrespective of its complexity. It achieves the main objective of object recognition that is to establish the identities of objects and use the information to predict future events. Taken together, maximal dependence capturing offers a single framework to achieve robust object representation and explain seemingly contradictory observations on the neurons' receptive field properties in different brain hierarchies.

MATERIALS AND METHODS

Learning Algorithm

Dictionary learning is treated as a blind source separation (BSS) problem (Comon, 1994; Comon and Jutten, 2010). An input signal is modeled as the response of M primary encoders. In the case of images, $M = m_1 \cdot m_2$, where m_1 and m_2 are the horizontal and vertical dimensions of the images. A set of N signals is presented for training as an input matrix $X \in \mathbb{R}^{M \times N}$, representing the response of M pixel to N patterns. The matrix X is then factorized into two matrices A and Φ, so that $X = \Phi A$. Here, $A \in \mathbb{R}^{K \times N}$ is the matrix representation of N patterns in a K dimensional basis set defined by $\Phi \in R^{M \times K}$.

To get the factor matrices through BSS, we imposed restriction on A to be sparse. The measure of sparsity was chosen to be l_0 norm, but the solution is achieved through minimizing l_1 norm. In addition to this sparsity constrain, we demanded both A and Φ to be non-negative.

Several possible BSS algorithms could result in an appropriate matrix decomposition under the given constraints (Hoyer, 2002, 2004; Rapin et al., 2013a; Allen et al., 2014). In particular, we used non-negative blind source separation algorithm nGMCA (Donoho and Elad, 2003; Rapin et al., 2013a,b). When a l_1 measure of sparseness is used, then the sum of the absolute values of coefficients of A is minimized. The minimization problem takes the form of:

$$\underset{\Phi,\mathbf{A}}{argmin}\ \frac{1}{2}||\mathbf{X}-\Phi\mathbf{A}||_2^2 + \lambda||\mathbf{A}||_1,\ subject\ to\ \mathbf{A} \geq 0;\ \ \Phi \geq 0$$

Thus, the process to solve this problem requires the minimization of the *Frobinius* norm difference (i.e., the Euclidean Distance) between the two sides of the equation and the minimization of the l_1 norm.

Each time BSS is performed, the Φ matrix was seeded with random numbers. Optimization was performed until convergence or when predefined number of iterations was reached.

Sparse Coding

Once the dictionary Φ is learned, any input pattern can be transformed into its corresponding representation. Transformation of input patterns is the process of finding the representation a that satisfies the equation: $\mathbf{x} = \Phi\mathbf{a}$. In our case, the dimension of the representational layer is chosen to be higher than that of the input layer, i.e., $K > M$. Here, decoding becomes an under-determined problem. Theories developed independently by Donoho (Donoho and Elad, 2003; Donoho, 2006a,b), and by Candes and Tao (Candes and Romberg, 2005; Candes et al., 2006; Candes and Tao, 2006) show that a unique solution can be obtained by imposing a sparseness constraint to the equation when solving the optimization problem. The most common use of sparsity definition includes l_0 and l_1. In our approach we perform l_1 minimization to solve:

$$\min ||\mathbf{a}||_1\ subject\ to\ ||\mathbf{x} - \Phi\mathbf{a}||_2 \leq \varepsilon$$

The l_1-minimization problem can be implemented by a standard convex optimization procedure, which can be found in several publications (Chen et al., 2001; Boyd and Vandenberghe, 2004; Candes and Tao, 2005; Donoho, 2006b; Donoho et al., 2012).

Redundancy Measurement

To measure redundancy in encoding objects, we treated the objects as following a uniform distribution, i.e., $\mathbb{P}(O_i) = 1/N$, where N is the total number of objects. The entropy of the ensemble of the objects is therefore $H(O) = logN$. We then calculated the capacity of the input

unit set (C) using the probabilities of occurrence of each encoder, $\mathbb{P}(x_i = 1) = p_i$:

$$C = \sum_{i=1}^{M} p_i log \frac{1}{p_i} + (1 - p_i) log \frac{1}{(1 - p_i)}$$

Redundancy was calculated as

$$R = 1 - \frac{H(O)}{C} = 1 - \frac{logN}{C}$$

The redundancy for representational units was calculated in a similar way, the only difference being that the representations were converted to binary forms using a Heaviside step function so that their l_0 norms could be considered while calculating probability of occurrence of individual encoders.

Kullback–Leibler Divergence Between Dictionary and Images

We used the Kullback–Leibler divergence (KL Divergence) to quantify the structural differences between symbols and dictionary elements. KL divergence [$D_{KL}(\mathbb{P}||\mathbb{Q})$] is a measure of information gained when a posterior probability distribution $\mathbb{P}$ is used to calculate the entropy instead of the prior distribution $\mathbb{Q}$. Denoting $\mathbb{Q}$ to be distribution over the states of a single pixel in symbol space and $\mathbb{P}$ to be the distribution over states of the same pixel in dictionary space, $D_{KL}(\mathbb{P}||\mathbb{Q})$ measures the information gained in considering the pixel to be coming from a dictionary element rather than symbols. A low divergence for all the pixels indicates that there is no gain in information if we consider any pixel to be coming from dictionary, indicating that the structure of the dictionary elements is same as structure of the symbols.

To calculate the distribution over the states of pixels in the dictionary space, all dictionary elements were binarized using a Heaviside step function. Probability of occurrence of individual pixels was calculated based on the number of dictionary elements in which the pixel is active. For instance, if a particular pixel x_i was active in n out of K dictionary elements, then the probability of occurrence of pixel x_i was calculated as

$$\mathbb{P}(x_i = 1) = \frac{n}{\mathbf{K}}$$

Probability of occurrence of the same pixel in symbol space is calculated based on the number of symbols m in which it is active i.e.,

$$\mathbb{Q}(x_i = 1) = \frac{m}{\mathbf{N}}$$

Here N is the number of symbols being encoded. Finally, the KL Divergence between the two distributions is calculated as

$$D_{KL}(x_i) = \mathbb{P}(x_i = 1) log \frac{\mathbb{P}(x_i = 1)}{\mathbb{Q}(x_i = 1)} + (1 - \mathbb{P}(x_i = 1)) log \frac{(1 - \mathbb{P}(x_i = 1))}{(1 - \mathbb{Q}(x_i = 1))}$$

Specificity Calculations

To quantify the specificity of a representational vector in representing the original object, we computed Z-scored similarity. Cosine similarity score between the representation of the test object (a_{test}) and all objects in the training set ($A_{training}$) were calculated and Z-scored. A high Z-score indicated high similarity between the representations of the test object and a particular object in $A_{training}$. In the figures, we plot the Z-scores for altered images with their unadulterated counterpart, which show high specificity in representing the original object.

Simulating Corrupted Signals

To test the robustness of object representation by the MDC framework, signals from the training set were selected and corrupted. The corrupted signals were subject to sparse decoding to generate their representations, which were then compared with those of the signals in the training set. We performed the following three types of corruption:

Noise-added corruption: we introduce noise by adding a Gaussian i.i.d. matrix $\mathcal{N}$ of varying standard deviation to the input matrix X. i.e., $X_{\mathcal{N}} = X + \mathcal{N}$, where $X_{\mathcal{N}} \in \mathbb{R}^{M \times N}$, is a matrix representation of noisy input. For Monte Carlo analysis, as described below, each time a simulation was performed, a different noise matrix $\mathcal{N}$ was introduced.

Pixel corruption: For a given signals, a fraction of the M pixels was selected from the input. Their values were maintained whereas the coefficients of the rest were set to zero.

Occlusion: For images, a contiguous set of pixels were selected, and their values were set to zero.

Monte Carlo Analysis

We performed Monte Carlo simulations by applying pixel corruption to the input signals and varying the number of corrupted pixels. 100 random sets (numbers varied from 2 to M) of pixels were selected. Using each of these randomly chosen sets, we performed sparse decoding to generate representation of the input patterns.

Input Identification

To calculate the correct identification of the object, we used the representation of each input in the training set as a library. The representation of each corrupted signal was compared with that in the library and cosine distances were computed. An input pattern was considered correctly identified if the cosine distances between its representation and that of the original signal was minimum (smaller than with representation of other patterns).

Projective Fields Generation From Natural Images

The mammalian visual systems process visual information in On and Off channels. On channel images were the normal images whereas Off channel images were the inverted images. To simulate parallel processing of the two channels, the On and Off images were concatenated along the rows and dictionary elements were generated by performing BSS on the concatenated matrix. The projective fields were constructed by superposing the On-channel portion of the dictionary element with the negative of Off-channel portion of the dictionary element.

Data

Symbols: A set of 1,000 symbols from word languages were obtained and digitized to 16 × 16 pixel arrays.

Natural images: Natural scenes from Van Hateren data base (Van Hateren and van der Schaaf, 1998) were digitized as grayscale pictures. Image patches of 16 × 16 size were randomly selected from the images. A total of 3,000 patches were used to form a training set.

Facial images: For face recognition, 2,000 frontal faces were obtained from Google search of publicly available images, trimmed and resized to 25 × 25 pixels. The Yale Face Database is obtained from http://cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

CRY conceived the idea and supervised the research. CRY and RR developed key concepts and co-wrote the manuscript. RR and DD performed analyses and modeling. KD generated face database. All authors contributed to the article and approved the submitted version.

FUNDING

The work is supported by funding from Stowers Institute and the NIH R01DC 014701.

ACKNOWLEDGMENTS

We would like to thank K. Si, J. Unruh, P. Kulesa, M. Klee and members of the Yu laboratory for insightful discussions. This work fulfills, in part, requirements for RR's and KD's thesis with the Open University, United Kingdom.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2022.857653/full#supplementary-material

REFERENCES

Allen, G. I., Grosenick, L., and Taylor, J. (2014). A generalized least-square matrix decomposition. *J. Am. Stat. Assoc.* 109, 145–159. doi: 10.1080/01621459.2013.852978

Anderson, C. H., and Vanessen, D. C. (1987). Shifter circuits: a computational strategy for dynamic aspects of visual processing. *Proc. Natl. Acad. Sci.* 84, 6297–6301. doi: 10.1073/pnas.84.17.6297

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Network* 3, 213–251. doi: 10.3109/0954898X.2011.638888

Atick, J. J., and Redlich, A. N. (1990). Towards a theory of early visual processing. *Neur. Comp.* 2, 308–320. doi: 10.1162/neco.1990.2.3.308

Attneave, F. (1954). Some informational aspects of visual perception. *Psychol. Rev.* 61, 183–193. doi: 10.1037/h0054663

Bao, P., She, L., McGill, M., and Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature* 583, 103–108. doi: 10.1038/s41586-020-2350-5

Barlow, H. (1961). "Possible principles underlying the transformation of sensory messages," in *Sensory communication*, ed. R. Wa (Cambridge, MA: MIT), 217–234.

Barlow, H. (1995). *The neuron in perception*. The cognitive neurosciences. Cambridge (MA): MIT Press.

Barlow, H. (2001). Redundancy reduction revisited. *Network* 12, 241–253. doi: 10.1080/net.12.3.241.253

Barlow, H. B. (1953). Summation and inhibition in the frog's retina. *J. Physiol.* 119, 69–88. doi: 10.1113/jphysiol.1953.sp004829

Barlow, H. B. (1989). Unsupervised learning. *Neur. Comp.* 1, 295–311.

Barlow, H. B., Fitzhugh, R., and Kuffler, S. (1957). Change of organization in the receptive fields of the cat's retina during dark adaptation. *J. Phys.* 137, 338–354. doi: 10.1113/jphysiol.1957.sp005817

Barwich, A.-S. (2014). A sense so rare: Measuring olfactory experiences and making a case for a process perspective on sensory perception. *Biol. Theory* 9, 258–268.

Barwich, A.-S. (2018). *Measuring the world: olfaction as a process model of perception*. Everything flows: Towards a processual philosophy of biology.

Barwich, A.-S. (2019). A critique of olfactory objects. *Front. Psychol.* 10:1337. doi: 10.3389/fpsyg.2019.01337

Bell, A. J., and Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Res.* 37, 3327–3338. doi: 10.1016/s0042-6989(97)00121-1

Bialek, W., Nemenman, I., and Tishby, N. (2001). Predictability, complexity, and learning. *Neur. Comp.* 13, 2409–2463. doi: 10.1162/089976601753195969

Biederman, I. (1985). Human image understanding: Recent research and a theory. *Comp. Vis. Graph. Image Proc.* 32, 29–73. doi: 10.1016/0734-189x(85)90002-7

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115–147. doi: 10.1037/0033-295x.94.2.115

Boussaoud, D., Desimone, R., and Ungerleider, L. G. (1991). Visual topography of area TEO in the macaque. *J. Comp. Neurol.* 306, 554–575. doi: 10.1002/cne.903060403

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience. *Psycholog. Rev.* 116:220. doi: 10.1037/a0014462

Boyd, S. P., and Vandenberghe, L. (2004). *Convex optimization*. Cambridge, MA: Cambridge University Press, 716.

Bruce, C., Desimone, R., and Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *J. Neurophys.* 46, 369–384. doi: 10.1152/jn.1981.46.2.369

Brunelli, R., and Poggio, T. (1993). Face recognition: Features versus templates. *IEEE Trans. Patt. Anal. Mach. Intell.* 15, 1042–1052. doi: 10.1109/34.254061

Buhmann, J., Lades, M., and von der Malsburg, C. (1990). "Size and distortion invariant object recognition by hierarchical graph matching," in *1990 IJCNN International Joint Conference on Neural Networks*. 1990, (IEEE).

Burge, T. (2010). *Origins of objectivity*. Oxford: Oxford University Press.

Burr, D. J. (1981). *Elastic matching of line drawings*. New York, NY: IEEE, 708–713.

Candes, E. J., Romberg, J. K., and Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* 59, 1207–1223. doi: 10.1002/cpa.20124

Candes, E. J., and Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Inform. Theory* 51, 4203–4215.

Candes, E. J., and Tao, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inform. Theory* 52, 5406–5425.

Candes, E. and Romberg, J. (2005). *l1-magic: Recovery of sparse signals via convex programming*. Available online at: URL: www.acm.caltech.edu/l1magic/downloads/l1magic.pdf. (accessed date April:14, 2005)

Chang, L., and Tsao, D. Y. (2017). The code for facial identity in the primate brain. *Cell* 169, 1013.–1028.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Rev.* 43, 129–159. doi: 10.1137/s003614450037906x

Chen, Y., Paiton, D. M., and Olshausen, B. A. (2018). The sparse manifold transform. Advances in neural information processing systems, 31.

Cireşan, D., Meier, U., Masci, J., and Schmidhuber, J. (2011). "A committee of neural networks for traffic sign classification," in *The 2011 international joint conference on neural networks*, (New York, NY: IEEE).

Cireşan, D. C., Meier, U., Masci, J., Gambardella, L. M., and Schmidhuber, J. (2011). "Flexible, high performance convolutional neural networks for image classification," in *Twenty-second international joint conference on artificial intelligence*, (Palo Alto, CA: AAAI Press).

Comon, P. (1994). Independent component analysis, a new concept? *Sign. Proc.* 36, 287–314.

Comon, P., and Jutten, C. (2010). *Handbook of blind source separation : independent component analysis and applications*. Amsterdam: Elsevier, 831.

Crawford, A., and Fettiplace, R. (1980). The frequency selectivity of auditory nerve fibres and hair cells in the cochlea of the turtle. *J. Phys.* 306, 79–125. doi: 10.1113/jphysiol.1980.sp013387

Crawford, A., and Fettiplace, R. (1981). An electrical tuning mechanism in turtle cochlear hair cells. *J. Phys.* 312, 377–412. doi: 10.1113/jphysiol.1981.sp013634

Damasio, A. R., and Damasio, H. (1993). "Cortical systems underlying knowledge retrieval: Evidence from human lesion studies," in *Exploring Brain Functions: Models in Neuroscience*, eds T. A. Poggio and D. A. Glaser (Hoboken, NJ: John Wiley and Sons), 233–233.

Damasio, A. R., Tranel, D., and Damasio, H. (1990). Face agnosia and the neural substrates of memory. *Annu. Rev. Neurosci.* 13, 89–109. doi: 10.1146/annurev.ne.13.030190.000513

Dan, Y., Atick, J. J., and Reid, R. C. (1996). Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *J. Neurosci.* 16, 3351–3362. doi: 10.1523/JNEUROSCI.16-10-03351.1996

Desimone, R., Albright, T. D., Gross, C. G., and Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.* 4, 2051–2062. doi: 10.1523/JNEUROSCI.04-08-02051.1984

DiCarlo, J. J., and Cox, D. D. (2007). Untangling invariant object recognition. *Trends Cogn. Sci.* 11, 333–341. doi: 10.1016/j.tics.2007.06.010

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron* 73, 415–434. doi: 10.1016/j.neuron.2012.01.010

Dong, D. W., and Atick, J. J. (1995). Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Comput. Neur. Syst.* 6, 159–178. doi: 10.1088/0954-898x_6_2_003

Donoho, D. L. (2006b). For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.* 59, 797–829. doi: 10.1002/cpa.20132

Donoho, D. L. (2006a). Compressed sensing. *Inform. Theory IEEE Trans.* 52, 1289–1306.

Donoho, D. L., and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via $\ell 1$ minimization. *Proc. Natl. Acad. Sci.* 100, 2197–2202. doi: 10.1073/pnas.0437847100

Donoho, D. L., Tsaig, Y., Drori, I., and Starck, J.-L. (2012). Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. *IEEE Trans. Inform. Theor.* 58, 1094–1121. doi: 10.1109/tit.2011.2173241

Fantana, A. L., Soucy, E. R., and Meister, M. (2008). Rat olfactory bulb mitral cells receive sparse glomerular inputs. *Neuron* 59, 802–814. doi: 10.1016/j.neuron.2008.07.039

Felleman, D. J., and Vanessen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cort.* 1, 1–47. doi: 10.1093/cercor/1.1.1-a

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Josa a* 4, 2379–2394. doi: 10.1364/josaa.4.002379

Field, D. J. (1994). What is the goal of sensory coding? *Neur. Comp.* 6, 559–601. doi: 10.1162/neco.1994.6.4.559

Fiser, J., and Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psycholog. Sci.* 12, 499–504. doi: 10.1111/1467-9280.00392

Fukushima, K. (2003). Restoring Partly Occluded Patterns: a Neural Network Model with Backward Paths. *Artif. Neur. Netw. Neur. Inform. Proc.* 2003, 393–400.

Fukushima, K. (2005). Restoring partly occluded patterns: a neural network model. *Neur. Netw.* 18, 33–43. doi: 10.1016/j.neunet.2004.05.001

Fukushima, K., and Miyake, S. (1982). "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," in *Competition and cooperation in neural nets*, eds S.-I. Amari and M. A. Arbib (Berlin: Springer), 267–285. doi: 10.1007/978-3-642-46466-9_18

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annu. Rev. Psychol.* 59, 167–192. doi: 10.1146/annurev.psych.58.110405.085632

Goldsmith, T. H. (1990). Optimization, constraint, and history in the evolution of eyes. *Q. Rev. Biol.* 65, 281–322. doi: 10.1086/416840

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT press.

Griffiths, T. D., and Warren, J. D. (2004). What is an auditory object? *Nat. Rev. Neurosci.* 5, 887–892. doi: 10.1038/nrn1538

Gross, C. G. (1992). Representation of visual stimuli in inferior temporal cortex. *Philosoph. Trans. R. Soc. Lond. Ser. B* 335, 3–10. doi: 10.1098/rstb.1992.0001

Gross, C. G. (2002). Genealogy of the "grandmother cell". *Neuroscientist* 8, 512–518. doi: 10.1177/107385802237175

Gross, C. G., Bender, D. B., and Rocha-miranda, C. E. (1969). Visual receptive fields of neurons in inferotemporal cortex of the monkey. *Science* 166, 1303–1306. doi: 10.1126/science.166.3910.1303

Hamilton, L. S., Oganian, Y., Hall, J., and Chang, E. F. (2021). Parallel and distributed encoding of speech across human auditory cortex. *Cell* 184, 4626–4639. doi: 10.1016/j.cell.2021.07.019

Hartline, H. K., and Ratliff, F. (1972). "Inhibitory interaction in the retina of Limulus," in *Physiology of Photoreceptor Organs*, Vol. 7, ed. M. G. F. Fuortes (Berlin: Springer), 381–447. doi: 10.1007/978-3-642-65340-7_11

Hoffman, D. D., and Richards, W. A. (1984). Parts of recognition. *Cognition* 18, 65–96.

Hoyer, P. O. (2002). "Non-negative sparse coding," in *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, (IEEE).

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.* 5, 1457–1469.

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154. doi: 10.1113/jphysiol.1962.sp006837

Hubel, D. H., and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Phys.* 195, 215–243. doi: 10.1113/jphysiol.1968.sp008455

Hummel, J. E., and Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychol. Rev.* 99:480. doi: 10.1037/0033-295x.99.3.480

Johnson, J. S., and Olshausen, B. A. (2005). The recognition of partially visible natural objects in the presence and absence of their occluders. *Vis. Res.* 45, 3262–3276. doi: 10.1016/j.visres.2005.06.007

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

Keller, A., Zhuang, H., Chi, Q., Vosshall, L. B., and Matsunami, H. (2007). Genetic variation in a human odorant receptor alters odour perception. *Nature* 449, 468–472. doi: 10.1038/nature06162

Kobayakawa, K., Kobayakawa, R., Matsumoto, H., Oka, Y., Imai, T., Ikawa, M., et al. (2007). Innate versus learned odour processing in the mouse olfactory bulb. *Nature* 450, 503–508. doi: 10.1038/nature06281

Konorski, J. (1967). *Integrative activity of the brain*. Chicago, IL: University of Chicago Press.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neur. Inform. Proc. Syst.* 25, 1097–1105.

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Ann. Mathem. Stat.* 22, 79–86.

Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung c* 36, 910–912. doi: 10.1515/znc-1981-9-1040

Laurent, G. (1999). A systems perspective on early olfactory coding. *Science* 286, 723–728. doi: 10.1126/science.286.5440.723

Lee, H., Ekanadham, C., and Ng, A. (2007). Sparse deep belief net model for visual area V2. *Adv. Neur. Inform. Proc. Syst.* 20, 873–880.

Lee, T. S., and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *JOSA A* 20, 1434–1448. doi: 10.1364/josaa.20.001434

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., and Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proc. IRE* 47, 1940–1951.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nat. Neurosci.* 5, 356–363. doi: 10.1038/nn831

Lindeberg, T. (2013). A computational theory of visual receptive fields. *Biol. Cyb.* 107, 589–635. doi: 10.1007/s00422-013-0569-z

Lindeberg, T. (2021). Normative theory of visual receptive fields. *Heliyon* 7:e05897. doi: 10.1016/j.heliyon.2021.e05897

Logothetis, N. K., Pauls, J., Bulthoff, H. H., and Poggio, T. (1994). View-dependent object recognition by monkeys. *Curr. Biol.* 4, 401–414. doi: 10.1016/s0960-9822(00)00089-0

Lörincz, A., Palotai, Z., and Szirtes, G. (2012). Efficient sparse coding in early sensory processing: lessons from signal recovery. *PLoS Comp. Biol.* 8:e1002372. doi: 10.1371/journal.pcbi.1002372

Ma, L., Qiu, Q., Gradwohl, S., Scott, A., Elden, Q. Y., Alexander, R., et al. (2012). Distributed representation of chemical features and tunotopic organization of glomeruli in the mouse olfactory bulb. *Proc. Natl. Acad. Sci.* 109, 5481–5486. doi: 10.1073/pnas.1117491109

Marr, D. (1969). A theory of cerebellar cortex. *J. Phys.* 202, 437–470. doi: 10.1113/jphysiol.1969.sp008820

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT press.

Marr, D., and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond. B* 200, 269–294. doi: 10.1098/rspb.1978.0020

Martinez, L. M., Wang, Q., Reid, R. C., Pillai, C., Alonso, M., Sommer, F. T., et al. (2005). Receptive field structure varies with layer in the primary visual cortex. *Nat. Neurosci.* 8, 372–379. doi: 10.1038/nn1404

McCulloch, W. S., Pitts, W., Lettvin, J., and Maturana, H. (1959). What the frog's eye tells the frog's brain. *Proc. IRE* 47, 1940–1951.

Meister, M., and Berry, M. J. (1999). The neural code of the retina. *Neuron* 22, 435–450. doi: 10.1016/s0896-6273(00)80700-x

Mombaerts, P. (2006). Axonal wiring in the mouse olfactory system. *Annu. Rev. Cell Dev. Biol.* 22, 713–737. doi: 10.1146/annurev.cellbio.21.012804.093915

Mombaerts, P., Wang, F., Dulac, C., Chao, S. K., Nemes, A., Mendelsohn, M., et al. (1996). Visualizing an olfactory sensory map. *Cell* 87, 675–686. doi: 10.1016/s0092-8674(00)81387-2

Niell, C. M., and Stryker, M. P. (2008). Highly selective receptive fields in mouse visual cortex. *J. Neurosci.* 28, 7520–7536. doi: 10.1523/JNEUROSCI.0623-08.2008

Nirenberg, S., Carcieri, S. M., Jacobs, A. L., and Latham, P. E. (2001). Retinal ganglion cells act largely as independent encoders. *Nature* 411, 698–701. doi: 10.1038/35079612

Ohiorhenuan, I. E., Mechler, F., Purpura, K. P., Schmid, A. M., Hu, Q., and Victor, J. D. (2010). Sparse coding and high-order correlations in fine-scale cortical networks. *Nature* 466, 617–621. doi: 10.1038/nature09178

Olshausen, B. A. (2013). *Highly overcomplete sparse coding*. in *Human Vision and Electronic Imaging XVIII*. Bellingham: International Society for Optics and Photonics.

Olshausen, B. A., Anderson, C. H., and Vanessen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.* 13, 4700–4719. doi: 10.1523/JNEUROSCI.13-11-04700.1993

Olshausen, B. A., and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609. doi: 10.1038/381607a0

Olshausen, B. A., and Field, D. J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vis. Res.* 37, 3311–3325. doi: 10.1016/s0042-6989(97)00169-7

Palm, G. (2013). Neural associative memories and sparse coding. *Neur. Netw.* 37, 165–171. doi: 10.1016/j.neunet.2012.08.013

Palmer, S. E., Marre, O., Berry, M. J., and Bialek, W. (2015). Predictive information in a sensory population. *Proc. Natl. Acad. Sci.* 112, 6908–6913. doi: 10.1073/pnas.1506855112

Pasupathy, A., and Connor, C. E. (2002). Population coding of shape in area V4. *Nat. Neurosci.* 5, 1332–1338. doi: 10.1038/972

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E., et al. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* 454, 995–999. doi: 10.1038/nature07140

Poggio, T., and Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature* 343, 263–266. doi: 10.1038/343263a0

Poggio, T., and Girosi, F. (1990). Networks for approximation and learning. *Proc. IEEE* 78, 1481–1497.

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., and Livingstone, M. S. (2019). Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. *Cell* 177, 999–1009e10. doi: 10.1016/j.cell.2019.04.005

Poo, C., and Isaacson, J. S. (2009). Odor representations in olfactory cortex:"sparse" coding, global inhibition, and oscillations. *Neuron* 62, 850–861. doi: 10.1016/j.neuron.2009.05.022

Puchalla, J. L., Schneidman, E., Harris, R. A., and Berry, M. J. (2005). Redundancy in the population code of the retina. *Neuron* 46, 493–504. doi: 10.1016/j.neuron.2005.03.026

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature* 435, 1102–1107. doi: 10.1038/nature03687

Rao, R. P. (1999). An optimal estimation approach to visual perception and learning. *Vis. Res.* 39, 1963–1989. doi: 10.1016/s0042-6989(98)00279-x

Rao, R. P., and Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neur. Comp.* 9, 721–763. doi: 10.1162/neco.1997.9.4.721

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Rapin, J., Bobin, J., Larue, A., and Starck, J.-L. (2013a). Sparse and non-negative BSS for noisy data. *Sign. Proc.IEEE Trans.* 61, 5620–5632. doi: 10.1109/tsp.2013.2279358

Rapin, J., Bobin, J., Larue, A., and Starck, L. (2013b) *Sparse Regularizations and Non-negativity in BSS*. Proceedings of SPARS, Lausanne, p. 83.

Reich, D. S., Mechler, F., and Victor, J. D. (2001). Independent and redundant information in nearby cortical neurons. *Science* 294, 2566–2568. doi: 10.1126/science.1065839

Ressler, K. J., Sullivan, S. L., and Buck, L. B. (1993). A zonal organization of odorant receptor gene expression in the olfactory epithelium. *Cell* 73, 597–609. doi: 10.1016/0092-8674(93)90145-g

Rey, H. G., Gori, B., Chaure, F. J., Collavini, S., Blenkmann, A. O., Seoane, P., et al. (2020). Single Neuron Coding of Identity in the Human Hippocampal Formation. *Curr. Biol.* 30, 1152.e–1159.e. doi: 10.1016/j.cub.2020.01.035

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.

Rieke, F., and Rudd, M. E. (2009). The challenges natural images pose for visual adaptation. *Neuron* 64, 605–616. doi: 10.1016/j.neuron.2009.11.028

Riesenhuber, M., and Poggio, T. (1999b). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2:1019. doi: 10.1038/14819

Riesenhuber, M., and Poggio, T. (1999a). Are cortical models really bound by the "binding problem"? *Neuron* 24, 87–93. doi: 10.1016/s0896-6273(00)80824-7

Riesenhuber, M., and Poggio, T. (2000). Models of object recognition. *Nat. Neurosci.* 3, 1199–1204.

Rodieck, R. W., and Rodieck, R. W. (1998). *The first steps in seeing*, Vol. 1. Sunderland, MA: Sinauer Associates.

Rolls, E. T. (1984). Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. *Hum. Neurobiol.* 3, 209–222.

Rolls, E. T. (2001). Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Vision* 2001, 366–395. doi: 10.1016/s0896-6273(00)00030-1

Rolls, E. T., and Milward, T. (2000). A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neur. Comp.* 12, 2547–2572. doi: 10.1162/089976600300014845

Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton (Project Para)*. New York, NY: Cornell Aeronautical Laboratory Inc, 29.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psycholog. Rev.* 65:386. doi: 10.1037/h0042519

Rozell, C. J., Johnson, D. H., Baraniuk, R. G., and Olshausen, B. A. (2008). Sparse coding via thresholding and local competition in neural circuits. *Neur. Comput.* 20, 2526–2563. doi: 10.1162/neco.2008.03-07-486

Rudd, M. E., Schwartz, G. W., and Rieke, F. (2009). Square-Root Law Light Adaptation in Rod-Mediated Vision is Due to Retinal Gain Control. *J. Vis.* 9, 89–89. doi: 10.1167/16.14.23

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *Internat. J. Comp. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Russell, I., and Sellick, P. (1978). Intracellular studies of hair cells in the mammalian cochlea. *J. Phys.* 284, 261–290. doi: 10.1113/jphysiol.1978.sp012540

Russell, I., and Sellick, P. (1983). Low-frequency characteristics of intracellularly recorded receptor potentials in guinea-pig cochlear hair cells. *J. Phys.* 338, 179–206. doi: 10.1113/jphysiol.1983.sp014668

Salisbury, J. M., and Palmer, S. E. (2016). Optimal prediction in the retina and natural motion statistics. *J. Stat. Phys.* 162, 1309–1323. doi: 10.1186/s12868-016-0283-6

Sermanet, P., and LeCun, Y. (2011). "Traffic sign recognition with multi-scale convolutional networks," in *The 2011 International Joint Conference on Neural Networks*, (New York, NY: IEEE).

Serre, T. (2014). Hierarchical Models of the Visual System. *Encyclop. Comp. Neurosci.* 6, 1–12. doi: 10.1007/978-1-4614-7320-6_345-1

Simoncelli, E. P., and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annu. Rev. Neurosci.* 24, 1193–1216. doi: 10.1146/annurev.neuro.24.1.1193

Singer, W. (1999). Neuronal synchrony: a versatile code review for the definition of relations. *Neuron* 24, 49–64.

Singer, Y., Teramoto, Y., Willmore, B. D., Schnupp, J. W., King, A. J., and Harper, N. S. (2018). Sensory cortex is optimized for prediction of future input. *Elife* 7:e31557. doi: 10.7554/eLife.31557

Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M., and Tolias, A. S. (2019). Engineering a less artificial intelligence. *Neuron* 103, 967–979. doi: 10.1016/j.neuron.2019.08.034

Smith, B. C. (2015). *The chemical senses*. Oxford: Oxford University Press.

Smith, E. C., and Lewicki, M. S. (2006). Efficient auditory coding. *Nature* 439, 978–982.

Srinivasan, M. V., Laughlin, S. B., and Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. Lond. Ser. B. Biol. Sci.* 216, 427–459. doi: 10.1098/rspb.1982.0085

Srivastava, N., Mansimov, E., and Salakhudinov, R. (2015). "Unsupervised learning of video representations using lstms," in *International conference on machine learning*, (PMLR).

Stettler, D. D., and Axel, R. (2009). Representations of odor in the piriform cortex. *Neuron* 63, 854–864. doi: 10.1016/j.neuron.2009.09.005

Stevens, C. F. (2018). Conserved features of the primate face code. *Proc. Natl. Acad. Sci.* 115, 584–588. doi: 10.1073/pnas.1716341115

Stockman, A., Langendörfer, M., Smithson, H. E., and Sharpe, L. T. (2006). Human cone light adaptation: from behavioral measurements to molecular mechanisms. *J. Vis.* 6, 5–5. doi: 10.1167/6.11.5

Tanaka, K. (1992). Inferotemporal cortex and higher visual functions. *Curr. Opin. Neurob.* 2, 502–505. doi: 10.1016/0959-4388(92)90187-p

Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science* 262, 685–688. doi: 10.1126/science.8235589

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19, 109–139. doi: 10.1146/annurev.ne.19.030196.000545

Tanaka, K., Saito, C., Fukada, Y., and Moriya, M. (1990). "Integration of form, texture, and color information in the inferotemporal cortex of the macaque," in *Vision, memory and the temporal lobe*, (Tokyo: Elsevier).

Tanaka, K., Saito, H.-A., Fukada, Y., and Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J. Neurophys.* 66, 170–189. doi: 10.1152/jn.1991.66.1.170

Treisman, A. (1999). Solutions to the binding problem: progress through controversy and convergence. *Neuron* 24, 105–125. doi: 10.1016/s0896-6273(00)80826-0

Treloar, H. B., Feinstein, P., Mombaerts, P., and Greer, C. A. (2002). Specificity of glomerular targeting by olfactory sensory axons. *J. Neurosci.* 22, 2469–2477. doi: 10.1523/JNEUROSCI.22-07-02469.2002

Tsao, D. Y., and Livingstone, M. S. (2008). Mechanisms of face perception. *Annu. Rev. Neurosci.* 31, 411–437. doi: 10.1146/annurev.neuro.30.051606.094238

Turk-Browne, N. B., Junge, J. A., and Scholl, B. J. (2005). The automaticity of visual statistical learning. *J. Exp. Psychol. Gen.* 134:552. doi: 10.1037/0096-3445.134.4.552

Ullman, S. (1996). *High-level vision: Object recognition and visual cognition*. A Bradford Book, Vol. 2. Cambridge, MA: MIT press.

Ullman, S. (1998). Three-dimensional object recognition based on the combination of views. *Cognition* 67, 21–44. doi: 10.1016/s0010-0277(98)00013-4

Ullman, S., and Basri, R. (1991). Recognition by linear combination of models. *IEEE Trans. Patt. Anal. Mach. Intell.* 13, 992–1006.

Ullman, S., Vidal-Naquet, M., and Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nat. Neurosci.* 5, 682–687. doi: 10.1038/nn870

Van den Bergh, G., Zhang, B., Arckens, L., and Chino, Y. M. (2010). Receptive-field properties of V1 and V2 neurons in mice and macaque monkeys. *J. Comp. Neurol.* 518, 2051–2070. doi: 10.1002/cne.22321

Van Hateren, J. H., and van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. Lond. Ser. B: Biolog. Sci.* 265, 359–366. doi: 10.1098/rspb.1998.0303

Vassar, R., Ngai, J., and Axel, R. (1993). Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium. *Cell* 74, 309–318. doi: 10.1016/0092-8674(93)90422-m

Von der Malsburg, C. (1995). Binding in models of perception and brain function. *Curr. Opin. Neurob.* 5, 520–526. doi: 10.1016/0959-4388(95)80014-x

Yamins, D. L., and DiCarlo, J. J. (2016b). Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365.

Yamins, D. L., and DiCarlo, J. J. (2016a). Eight open questions in the computational modeling of higher sensory cortex. *Curr. Opin. Neurobiol.* 37, 114–120. doi: 10.1016/j.conb.2016.02.001

Young, M. P., and Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science* 256, 1327–1331. doi: 10.1126/science.1598577

Yuille, A. L. (1991). Deformable templates for face recognition. *J. Cogn. Neurosci.* 3, 59–70. doi: 10.1162/jocn.1991.3.1.59

Conflict of Interest: RR, DD, and CRY declare the existence of a financial competing interest in the form of a patent application based on this work.

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

Publisher's Note: 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.