ANNUAL REVIEWS

Annual Review of Psychology

Understanding Human Object Vision: A Picture Is Worth a Thousand Representations

Stefania Bracci[1] and Hans P. Op de Beeck[2]

[1]Center for Mind/Brain Sciences, University of Trento, Rovereto, Italy; email: stefania.bracci@unitn.it

[2]Leuven Brain Institute, Research Unit Brain & Cognition, KU Leuven, Leuven, Belgium; email: hans.opdebeeck@kuleuven.be

Annu. Rev. Psychol. 2023. 74:113–35

First published as a Review in Advance on November 15, 2022

The *Annual Review of Psychology* is online at psych.annualreviews.org

https://doi.org/10.1146/annurev-psych-032720-041031

Keywords

object recognition, visual cortex, object representations, deep convolutional neural networks, DCNNs, behavior

Abstract

Objects are the core meaningful elements in our visual environment. Classic theories of object vision focus upon object recognition and are elegant and simple. Some of their proposals still stand, yet the simplicity is gone. Recent evolutions in behavioral paradigms, neuroscientific methods, and computational modeling have allowed vision scientists to uncover the complexity of the multidimensional representational space that underlies object vision. We review these findings and propose that the key to understanding this complexity is to relate object vision to the full repertoire of behavioral goals that underlie human behavior, running far beyond object recognition. There might be no such thing as core object recognition, and if it exists, then its importance is more limited than traditionally thought.

Contents

1. INTRODUCTION

The recognition of objects has long been considered a relatively unitary goal of the visual system. Together with the underlying neural pathway, referred to as the ventral visual "what" pathway, object recognition, as the major component of object perception, is usually contrasted with the "where/how" dorsal pathway computations that evolved for the purpose of more spatial and action-related tasks (e.g., Goodale & Milner 1992). Based on this assumption, research in object vision has attempted to understand how the ventral pathway solves the object recognition problem and what code is used to represent the countless objects we constantly, and effortlessly, perceive from the moment we open our eyes. Here, we argue that to better understand human object vision, we need to consider object recognition within the context of the human's unique and rich cognitive and behavioral repertoire. Objects and their features are represented for the purpose of many goals, and this is reflected in the presence of very rich representations.

As a real-life example, think of the following. On your daily commute to work, you encounter many objects such as buildings, cars, traffic lights, people, and so on. Depending on a given goal, your brain extracts differential information from these objects. Some objects are likely to be exploited as landmarks during navigation (e.g., crossroads, buildings, streets), and others are to be processed for their functional and action-related properties (e.g., your office key, the elevator button, the university's badge). Surely, your brain needs to recognize your office's key in order to exploit useful information from it, but the type of information extracted from the key (i.e., action related) is different from the type of information extracted from a busy crossroad (i.e., navigation). In addition, many objects might serve multiple goals depending on the context. Since the visual system does not just recognize objects for the sake of it, we need to consider how object representations can ultimately support behavioral goals in order to understand the object representational code. In other words, object recognition is not a unitary problem that we can hope to solve with a set of general properties equally useful in characterizing all objects.

This proposal is also supported by most recent developments in computational neuroscience. Deep convolutional neural networks (DCNNs) have impressed the worldwide community with their ability to reach human-level image classification behavior, yet they fail to explain the complexity of human visual cortex organization. Despite initial evidence suggesting interesting similarities in DCNNs' ability to capture the large-scale object space observed in the brain, further studies are now uncovering that large gaps still remain. We provide evidence suggesting that our framework extends to computational neuroscience and has the potential to guide the creation of DCNNs that might better capture human object vision. Specifically, the current computational gap between brain and DCNN representations can be interpreted in terms of training goals and scope.

In Section 2, we first give an overview of how decades of research have considered object recognition the ultimate computational goal of the ventral visual pathway and how this view has influenced the parallel progress of computational modeling in object vision. In Section 3, we go on to describe the diversity of content that characterizes object representations, highlighting the relevance of multiple and overlapping dimensions that span from low-level visual properties to high-level semantic attributes. In Section 4, we lay out further diversity in the ventral pathway representational space explained by embedding object recognition in the larger context of functional behavior. We progress in parallel by alternating between a cognitive and computational focus, which are necessarily interrelated and back up similar conclusions.

2. OBJECT RECOGNITION AND THE VISUAL VENTRAL PATHWAY

The question of how we can effortlessly recognize myriads of objects within a few milliseconds, with high accuracy, and despite the huge variation of input coming from our eyes, is a fascinating one that has attracted the attention of many brain scientists. Initial psychological attempts to understand object perception integrated behavioral findings with quantitative approaches (e.g., Attneave 1957, Shepard & Chipman 1970), but since the seminal discoveries made by Hubel & Wiesel (1968) for simple and complex cells, cognitive neuroscience and computational neuroscience have become increasingly important for the scientific understanding in this domain. Here, we provide a general summary of the progress made in the last decades in this interdisciplinary endeavor.

From early on, object recognition has been the focus of investigations in visual information processing, and there have been elegant yet relatively narrow views on the type of representation that needs to be built to achieve this goal (e.g., Marr & Nishihara 1978, Hoffman & Richards 1984, Biederman 1987). Marr (1980) described how this information processing would culminate in a 3D model representation as an object-centered description of the three-dimensional structure and organization of a viewed shape. While recent overviews consider much more elaborate underlying representations, the focus upon object recognition has remained—sometimes rephrased as core object recognition (DiCarlo et al. 2012).

The ventral visual pathway plays a critical role in object perception, as shown by seminal research in neuropsychology and in neurophysiology (Mishkin & Ungerleider 1982, Goodale & Milner 1992). Throughout the ventral stream's hierarchical processing in monkeys, the responses of neurons become increasingly selective from mid-level to high-level features; neurons in V4 respond to form, texture, and color (Desimone et al. 1984, Hu et al. 2020) and more anteriorly, in inferior temporal (IT) cortex, to object identity (Gross et al. 1972, Hung et al. 2005). Early on, neurons with very different selectivity seemed mostly intermingled in IT with little clustering (Baylis et al. 1987). Yet, neuropsychological case studies showed that representations of different object categories can be dissociated neuroanatomically. Focal brain lesions within ventral visual cortex cause category-specific recognition deficits (Caramazza & Shelton 1998) that cannot be accounted for by visual differences in the stimuli such as image complexity and/or familiarity with certain categories (Laiacona et al. 1993). More recently, transcranial magnetic stimulation confirmed that temporarily disrupting localized areas in visual cortex results in transient category-specific recognition deficits (Pitcher et al. 2009).

Neuroimaging results do also point to a computational progression from processing low/mid-level features relevant to supporting general object recognition to neural substrates devoted to elaborating information on specific object categories (**Figure 1**). The lateral occipitotemporal complex (LOC) appears to encode object shape regardless of category; it responds equally well to (*a*) an object, whether it is a 3D image or a 2D line drawing (Kourtzi & Kanwisher 2001);

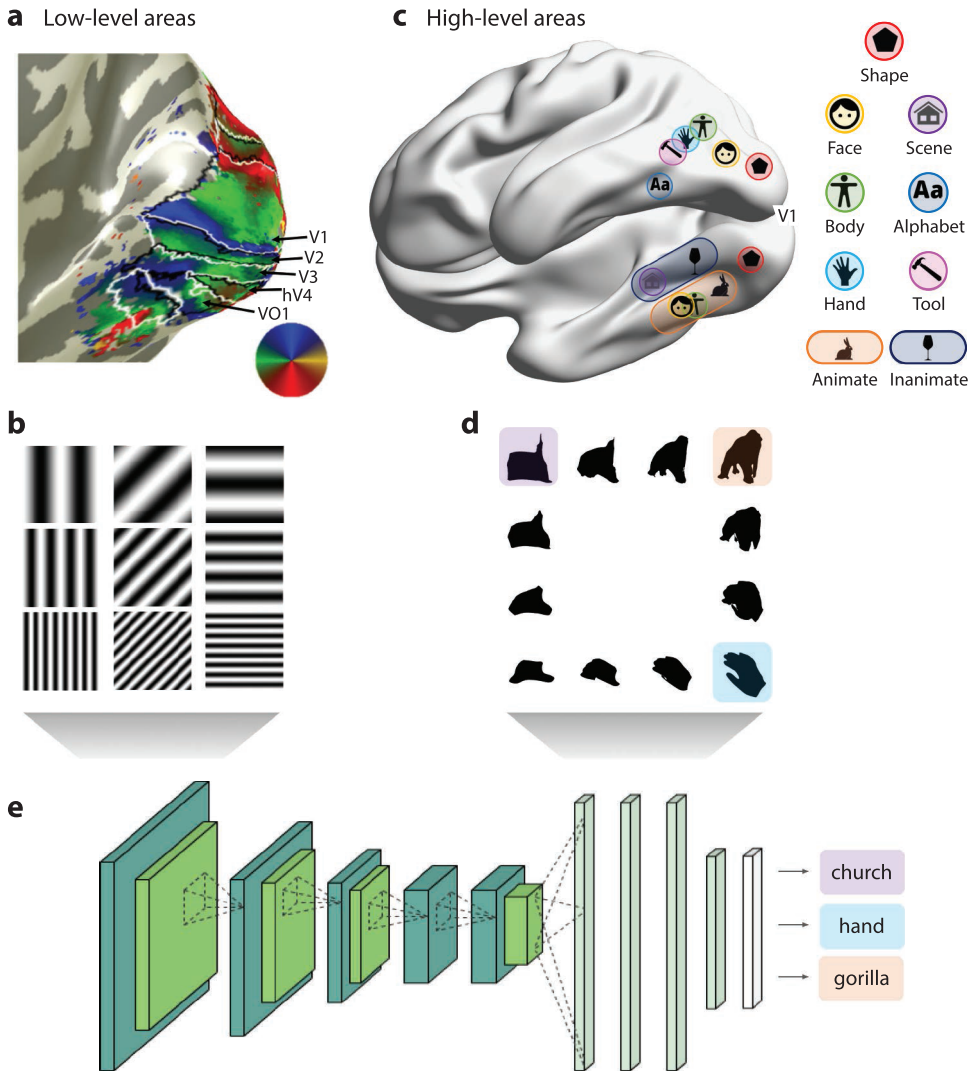

Figure 1

Visual processing hierarchy as emphasized in traditional models of core object recognition. There is a progression from processing of basic visual dimensions at lower levels (panels *a*, *b*) to a representation of object category and meaningful object dimensions at higher levels (panels *c*, *d*). Basic visual dimensions include (*a*) retinotopic position, which is the main organizing feature of visual areas V1 to V4 (image shows medial view of posterior occipital pole) as well as (*b*) dimensions such as orientation and spatial frequency. (*c*) High-level areas show strong focal selectivity for categorical distinctions (image shows the typical locations of selectivity areas, indicated with pictograms) and (*d*) contain multidimensional representations of objects (image shows a morphing space constructed with four objects: church, gorilla, hand, and hat). (*e*) A similar progression exists in deep convolutional neural networks that are trained in complex visual tasks such as image classification. Teal and green boxes on the left represent typical convolutional processing steps (respectively convolution and max pooling), which feed into fully connected layers toward the right. Dotted lines illustrate the interlayer connectivity. Panel *a* adapted from Gomez et al. (2018) (CC BY 4.0). Panel *b* adapted from **http://apps.usd.edu/coglab/schieber/images/sf_poster_3x6.jpg** with permission from Frank Schieber.

(*b*) the whole object as well as its parts (Grill-Spector et al. 1998); and (*c*) familiar and unfamiliar object shapes (Malach et al. 1995). Around and partially overlapping with LOC, occupying a large part of lateral and ventral occipitotemporal cortex (OTC), there are multiple category-selective areas, providing a confirmation of the predictions from earlier neuropsychological work. However, category selectivity has been reported only for a limited number of categories, including faces (Kanwisher et al. 1997), bodies (Downing et al. 2001), scenes (Epstein & Kanwisher 1998), tools (Chao et al. 1999), hands (Bracci et al. 2010), and letter strings (Cohen et al. 2002). Each of these categories is associated with multiple category-selective regions (Rosenke et al. 2021). For other object categories (cars, flowers, etc.), there are no regions identified with strong object preference. This does not directly translate to lack of selectivity for other object categories. Single neurons show selectivity for relatively fine object changes (e.g., Kayaert et al. 2005), and even in fMRI many different categories can be differentiated using sensitive analysis methods such as multivoxel pattern analyses (Cox & Savoy 2003). Yet, the selectivity and clustering seem particularly strong for a small subset of categories. Further developments in spatial resolution of brain imaging (e.g., scanning at higher magnetic fields), in analysis methods, and in the development of large-scale data sets containing the neural responses to thousands of images will very likely increase the number of categories and stimulus distinctions for which selectivity is found. This has recently been illustrated by the presence of selectivity for food in OTC (Jain et al. 2022, Khosla et al. 2022).

Quantitatively speaking, the dominance of a few categories can be understood when considering core object recognition as the one and only computational process at work. Strong category selectivity might reflect exceptional needs for within-category discrimination. Rosch et al. (1976) suggested that most objects are first recognized at the basic level (e.g., cats, dogs, persons). Biederman (1987) estimated that there might be around 1,500–3,000 basic-level categories, and that most of these categories might only contain a few discriminated types or exemplars. We can compare these numbers with the recent estimate that people typically know around 5,000 faces (Jenkins et al. 2018). Given these numbers, it might be reasonable that the neural territory devoted just to the domain of faces would be comparable to the resources shared by thousands of other object categories. However, the same reasoning might not work for other domains. In particular, it seems unlikely that we would specifically recognize thousands of bodies or hands.

These general characteristics in neural information processing in the human brain have been modeled in DCNNs (**Figure 1**). A very relevant early model is the HMAX standard model (Riesenhuber & Poggio 1999), which proposed an alternation of linear and nonlinear pooling mechanisms as an extension of the simple and complex cell scheme proposed by Hubel & Wiesel (1968). The HMAX model succeeded in capturing many of the basic phenomena found in electrophysiological studies in monkeys, such as tolerance for changes in object size and a graded tolerance for the viewpoint from which an object is seen. Yet, it failed to simulate more complex and fine-grained aspects of object vision, including the most prominent features and dimensions that determine whether humans consider two shapes as representing the same object or as being similar or not (Kayaert et al. 2005, Op de Beeck et al. 2008c)—properties that are often referred to as representational geometry.

In the last decade, deeper models (i.e., models with more hierarchical layers) have been developed. The typical approach is to train DCNNs on image classification with more than a million images that cover one thousand classes (ImageNet; Deng et al. 2009), a task on which DCNNs have outperformed other computer vision approaches since 2012 (Krizhevsky et al. 2012). A thousand categories is a decent number if compared with the aforementioned estimate obtained for humans, and it might be better than the knowledge scope in the most advanced nonhuman primate model: monkeys raised and tested in captivity.

With such networks, computational neuroscientists can mimic the major principles of information processing and how representations are transformed along the visual pathway starting from a 2D image. Initial convolutional layers contain filters that analyze the image or the previous layer responses locally, and the last layers are typically fully connected so that a unit has access to the responses of all units from the previous layer. This arrangement is conceptually similar to the increase in receptive field size that characterizes the successive processing stages in the human visual system (for review, see LeCun et al. 2015).

The information processing in DCNNs trained on object recognition corresponds to some degree with processing in the human and primate brain. In contrast to earlier networks, the representational geometry in these DCNNs overlaps to some extent with how humans perceive and represent objects and their shape (Kubilius et al. 2016). Representations in early convolutional layers correspond best with retinotopic visual areas, whereas fully connected layers capture important aspects of representations in higher-level areas in lateral and ventral occipitotemporal cortex (Güçlü & van Gerven 2015, Kar et al. 2019, Schrimpf et al. 2020, Lindsay 2021). Yet, the correspondences are not perfect, and the models can be fooled with adversarial images that would have no or at least less of an effect for human observers (Nguyen et al. 2015, Dujmović et al. 2020). Several approaches have been shown to further improve the correspondence between the representations in DCNNs and human vision, including (*a*) the addition of human-like limitations, such as dealing with noisy computations or an initial restriction to seeing only low spatial frequencies (e.g., Kim et al. 2020, Avberšek et al. 2021); (*b*) a specific focus on global shape as opposed to local features or texture (Geirhos et al. 2018, Baker et al. 2020); (*c*) the addition of feedback connections (Kietzmann et al. 2019); and (*d*) training with more ecologically valid image databases such as Ecoset (Mehrer et al. 2021). Nevertheless, the effect size of many of these improvements seems relatively small.

The traditional DCNN architecture and training regime does not provide sufficient information about why we would see the emergence of category-selective regions where neurons with similar preferences cluster. In neuroscience it is often proposed that clustering occurs because it helps minimizing the wiring cost of intracortical connectivity (Chklovskii & Koulakov 2004). When topographic DCNNs with such constraints are trained in image classification, they develop category-selective regions (Lee et al. 2020, Keller et al. 2021, Blauch et al. 2022). Even without such constraints, more traditional DCNNs show separated, module-like subsystems for face and object processing when they are trained in both domains (Dobs et al. 2022). Normally, DCNNs are trained for object classification, thus suggesting that the need for core object recognition might be sufficient to explain many of the properties of the human visual system. It does not seem necessary, though, because internal representations emerging in neural networks trained to generate naturally looking images or in networks trained in an unsupervised manner also tend to represent visually dissimilar images of the same category as being similar (Konkle & Alvarez 2022).

3. HOW ARE OBJECTS REPRESENTED IN THE VENTRAL VISUAL PATHWAY?

The previous section suggests that objects are represented in terms of high-level dimensions and that useful high-level representations would abstract away from low-level feature selectivity as it exists in lower processing stages, which was an assumption in classic theories of object recognition (Marr & Nishihara 1978, Biederman 1987). Experimental studies in monkeys motivated by these theories emphasized that IT neurons are not very sensitive to low-level dimensions such as stimulus position and size (e.g., Ito et al. 1995), and similar conclusions were drawn from human fMRI (Grill-Spector et al. 1999). However, this situation changed in the 2000s, starting with

several studies showing that IT neurons have surprisingly small receptive fields when probed with relatively small stimuli (Op de Beeck & Vogels 2000, DiCarlo & Maunsell 2003), and object representations in object-selective cortex are constrained by their position in the visual field (Kravitz et al. 2013).

Since then, evidence has come around for selectivity in OTC for a wide range of features that spans the full range, from low-level features typically associated with primary visual cortex all the way up to semantic object properties (**Figure 2**). Studies in the literature have varied in how much they emphasized the low-level selectivity versus high-level properties, and in some cases they argued that a particular level of selectivity is most important or is the primary factor from which

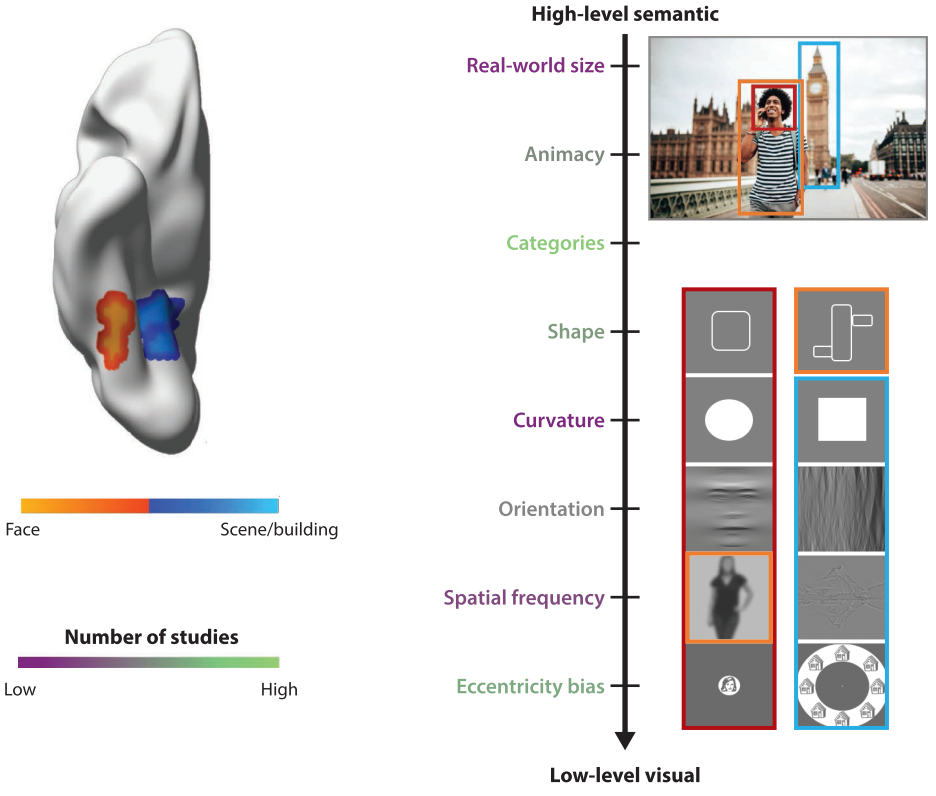

Figure 2

Higher levels of processing show selectivity for a wide variety of object and image properties, ranging from basic visual dimensions to semantic dimensions. The activity map on the left visualizes one of the most prominent categorical distinctions made in high-level visual cortex: faces versus scenes or buildings. The diagram on the right illustrates the range of features for which these areas have shown selectivity, with basic visual features at the bottom and semantic properties at the top. The bottom-up order of the features is based upon the first level of processing with which that feature is typically associated, starting at the retina for eccentricity bias and spatial frequency (adapted from Hasson et al. 2002, Canário et al. 2016, respectively, with permission from Elsevier), V1 for orientation [adapted from Goffaux & Dakin 2010 (CC BY 4.0)], V2-V4 for curvature, V4-LOC for shape, and lateral and ventral OTC for category, animacy, and real-world size (Konkle & Oliva 2012). The purple-to-green color-coded label of each feature is an ordinal indication of how many studies are found in a literature search aimed at studies of high-level areas (*purple*, a few; *gray*, a few hundred; *green*, thousands). Red, blue, and orange boxes represent the feature selectivity of face-, scene/building-, and body-selective regions, respectively. The top right image is adapted from **https://www.gettyimages.it/detail/foto/tourist-in-london-immagine-royalty-free/1285355810**.

other selectivity emerges. For example, it has been suggested that the existence of an eccentricity organization early in development, together with the tendency to foveate faces and words but not scenes, has caused face, word, and scene selectivity to emerge in specific anatomical locations (Hasson et al. 2002, Arcaro & Livingstone 2021). Moving all the way up to semantic accounts, other scholars have proposed that the organization of high-level visual cortex is driven by connectivity with domain-specific networks for social cognition (faces), language (words), and navigation (scenes and landmarks) (Mahon & Caramazza 2011, Saygin et al. 2012, Peelen & Downing 2017, Powell et al. 2018). It is not necessary to consider these options as being mutually exclusive, and multiple factors might be at work together (Op de Beeck et al. 2019).

To test the contribution of object visual properties in OTC object space, one approach has been to remove the influence of category information by presenting the object's structural properties only. Results from these studies (see **Figure 2** for illustration) show coherent feature maps that nicely correlate with category responses throughout OTC (Rajimehr et al. 2011, Nasr et al. 2014). Face selectivity overlaps with a preference for low spatial frequencies (SF) and curvilinear feature maps, whereas scene selectivity overlaps with high SF and rectilinear feature maps. Face- and scene-selective regions are also different in terms of mid-level feature selectivity, such as texforms (texture synthetized images that preserve texture information but cannot be recognized at the category level) (Long et al. 2018). Likewise, Jagadeesh & Gardner (2022) show that information supporting object categorization in OTC is accessible, but its representation appears to lack sensitivity to the spatial arrangement of features that characterize a specific object, thus suggesting a representation of complex texture-like properties. These findings add to the many studies showing that eccentricity maps that characterize low-level visual cortex for central versus peripheral vision can predict the anatomical location of category-selective representations in OTC (Levy et al. 2001, Hasson et al. 2002). The common denominator for exponents of visual accounts relies on the assumption that there is nothing special about category-selective representations in themselves. OTC is a general-purpose machine to recognize and categorize all objects; hence, category-related effects observed in OTC are simply the tip of an iceberg whose underlying neural geology can be better explained by distributed feature maps—of low and mid-level features—that span across the full OTC territory (Levy et al. 2001, Malach et al. 2002, Rajimehr et al. 2011, Baldassi et al. 2013, Nasr et al. 2014, Arcaro & Livingstone 2021). One possible criticism is that for some of the feature biases it has not been established that the response strength and amount of selectivity are equally strong or close to the response strength and selectivity at the category level. Yet, if we combine selectivity for multiple features together, it is conceivable that together they might explain the strength of selectivity at the category level even without the need to invoke nonlinear mechanisms that further amplify category selectivity (Op de Beeck et al. 2008b). For example, Hasson and colleagues (2002) already showed that the eccentricity preference in face- and scene-selective cortex might be as strong as the category selectivity.

There is a problem, though, with such a low-level account when we consider the combinations of feature biases. That is, one has to consider whether this account explains the exact combinations of preferred features. Such a description can work if the combination of feature preferences is inherited from early areas like V1. In some cases, it might be. For example, Arcaro & Livingstone (2021) suggested that a preference for curvature might go together with foveal receptive fields, even in early visual areas. However, in other cases this does not work. In early vision, foveal responses go together with a preference for higher spatial frequencies (Arcaro & Livingstone 2017). The receptor characteristics and density in the retina preclude a sensitivity for higher spatial frequencies at higher eccentricity. However, in high-level visual cortex, the foveal bias in face-selective cortex goes together with a preference for lower spatial frequencies, and vice versa in scene-selective cortex (i.e., peripheral bias and preference for higher spatial frequencies;

Canário et al. 2016). Such peculiar feature preferences suggest that there is another reason we see these combined feature preferences. Bracci and colleagues (2017) suggested that this reason can be found at the level of category coding. In this framework, at least some of the feature maps of object structural properties observed in OTC might be a result of category specificity/preference rather than its cause. In other words, through visual experience, category-selective areas develop feature-specific biases due to the lifelong exposure to natural co-occurrences of an object (e.g., face) and its structural properties (e.g., oval shape). Note that this assertion does not exclude the possibility that some of the feature biases, such as eccentricity, might have played a role in driving where selectivity clusters have emerged.

While it is worth paying attention to the evidence for feature biases, we should not forget that most studies of high-level visual cortex have focused upon the category level (see green text in the right side of **Figure 2**). This is how the functional architecture of lateral and ventral OTC was first defined, with areas selective for specific categories. Evidence for category-selective deficits following focal lesions in OTC (Barton et al. 2002) and fMRI evidence for category selectivity and behavioral effects of object recognition (Yovel & Kanwisher 2005) do speak in favor of a strong role for categorical coding. The category-level organization is not just one level, though. The object space in OTC can be understood as a hierarchical semantic space. At the superordinate level, the object space in OTC is separated along the mid-fusiform sulcus: The lateral fusiform gyrus (FG) represents animate entities, and the medial FG represents inanimate entities. A strong role of animacy has been confirmed in multiple studies. Within this macro division, at the ordinate level different islands of selectivity can be found for animals, faces, and bodies within the animate domain and for objects, tools, and places within the inanimate domain (e.g., Kriegeskorte et al. 2008, Grill-Spector & Weiner 2014).

Several studies have tried to disentangle the role of shape and semantic properties for this superordinate categorical dimension of animacy. Shape and semantics are highly correlated in our visual experience of objects. Faces are round and characterized by a unique configuration of features; scenes are rich in rectilinear shapes. Attempts at orthogonalizing object shape and semantic properties have revealed the importance of both dimensions (Bracci & Op de Beeck 2016, Proklova et al. 2016). A recent study in nonhuman primates confirmed the contribution of both dimensions and more specifically defined the most dominant visual dimension (Bao et al. 2020), summarizing the first two dimensions of object space by category information (animate–inanimate) and object aspect ratio (stubby–spiky). The role of aspect ratio fits with earlier monkey fMRI work showing selective regions for stubby and spiky objects (Op de Beeck et al. 2008a). The conceptualization in terms of these two dimensions is also in line with previous reports pointing to continuous OTC maps for semantic categories (e.g., animacy continuum) as well as for aspect ratio of objects (Kriegeskorte et al. 2008, Op de Beeck et al. 2008c, Connolly et al. 2012, Baldassi et al. 2013, Sha et al. 2015) or shape more in general (Bracci & Op de Beeck 2016). Together, these results point to an important role played by these dimensions in OTC object space (**Figure 3**). Note that **Figure 3** shows the best examples for this proposal; in other cases only one of these dimensions is seen, such as only aspect ratio (Baldassi et al. 2013) or only animacy (human fMRI data in Kriegeskorte et al. 2008).

However, it would be a mistake to believe that any small set of dimensions would provide a good characterization of object representations, be it aspect ratio (or shape in general) and animacy or any other set. Useful object representations show tuning for a rich repertoire of features and dimensions, including low-level feature biases, many shape properties, and a multidimensional categorical space. Finding a few primary dimensions is often a consequence of a study design that neglects certain dimensions. As a simple example, the proposal in terms of two object dimensions, animacy and aspect ratio, came from a study that showed all stimuli at the fovea (Bao et al. 2020). If

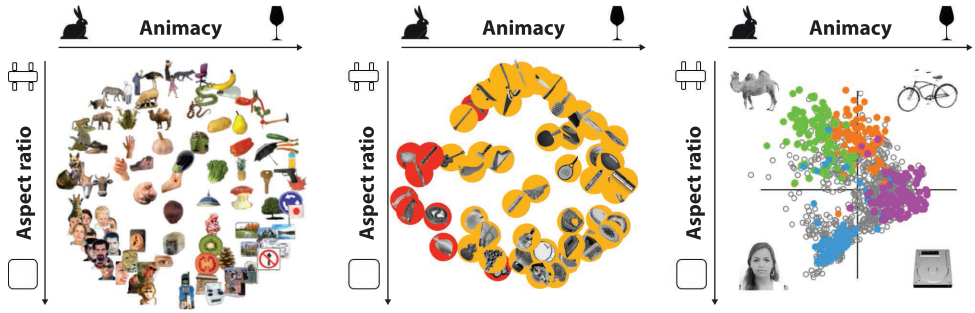

Figure 3

Out of all object properties that are represented in primate occipitotemporal cortex, two dimensions seem particularly important: animacy and aspect ratio. This finding was observed in studies with very different designs and methods, including (*left*) monkey single-cell recordings with a broad set of object images (adapted from Kriegeskorte et al. 2008 with permission from Elsevier), (*middle*) human fMRI with a design that explicitly dissociates shape from category membership (adapted from Bracci & Op de Beeck 2016), and (*right*) monkey fMRI and single-unit recordings (adapted from Bao et al. 2020 with permission from Springer). The two dimensions are plotted so that animacy decreases from left to right, and aspect ratio [as defined by Bao et al. (2020)] decreases from top to bottom.

object position had been varied, then position would also have been an important factor. Thus, object representations are much higher-dimensional than one would deduce from such experiments.

The relevance of high-dimensional representations with feature selectivity at multiple levels is confirmed by computational modeling with DCNNs. Hong and colleagues (2016) investigated properties such as position, size, and pose, and they showed that both monkey inferior temporal population responses and DCNNs show an increase along the processing hierarchy in terms of how easily such properties can be decoded. Population activity in inferior temporal cortex (or a fully connected DCNN layer) is more useful than activity in primary visual cortex (or a convolutional layer) to determine where an object is. This might seem counterintuitive, but note that the coding of position here is not only the ability to say that something is present in an otherwise empty scene, for which V1 responses would be perfectly fine, but also the ability to localize an object in a complex and cluttered scene.

Li & Bonner (2021) trained a classifier to respond selectively to scenes using as input the responses of a late convolutional layer, a layer that arguably represents features of a moderate complexity. This approach can be seen as a model of how category selectivity (in this case, scene selectivity) can emerge by taking responses from a nonselective filter bank like V4. Interestingly, the resulting classifier not only was selective for other scenes but also showed a complex profile of selectivity for many scene-relevant features that is reminiscent of what has been demonstrated for the human parahippocampal place area.

Likewise, Zeman et al. (2020) showed that fully connected DCNN layers show a joint tuning for shape and category in human OTC, as found in the human brain by Bracci & Op de Beeck (2016). Also, in this case, similar to the findings of Hong and colleagues (2016), the tuning of more complex aspects of shape actually increases in the brain regions that show clear category selectivity. The better a representation is for object recognition, the richer it is. This assertion suggests that there are many reasons to believe that object representational spaces are high-dimensional. When low-dimensional spaces are reported, results are often caused by using pure noise (useless) representations as a baseline and by working with a relatively small and very ordered stimulus space. In the latter case, a useful internal representation should indeed also be low-dimensional

(e.g., Op de Beeck et al. 2001), and revealing such a regularization in biological and modeled brains is important. However, it is the regularization that is the key point in such studies, not the low dimensionality. The full object space turns out to be very high-dimensional in addition to being very ordered, at least when probed with sufficiently elaborate stimulus designs. For example, Morgenstern and colleagues (2021) used over 100 shape features and a model with 22 linear combinations of these features to characterize just one aspect of objects, their perceived shape. Whereas aspect ratio was (again) present as an obvious dimension in a low-dimensional plot of the shape space, a model with just one or a few dimensions performed poorly. Likewise, Hebart and colleagues (2020) identified 49 highly reproducible object dimensions that explain most variance in human similarity judgments. Dimensions varied substantially in complexity, from labels such as "colorful" and "round" up to "animal-related" and "eating-related." As a computational argument, Elmoznino & Bonner (2022) observed that DCNNs with a higher dimensional representation show better generalization performance to previously unseen stimuli.

What could be the reason for an object space with such a high dimensionality? At first sight this might be counterintuitive. An object code that allows fast object recognition might be more efficient if all objects could be described by a general set of dimensions. However, object recognition does not happen out of context. Therefore, to understand the code used by the brain to represent objects, we need to consider object recognition in the context of higher-level behaviorally relevant computations. We address this point in the next section.

4. A PICTURE IS WORTH A THOUSAND REPRESENTATIONS

Our visual behavior is guided by our need to move in and interact with our environment. Thus, representations that support object recognition might be entangled with higher, and more comprehensive, domain-specific representations that support our actions and spatial and social behavior. We suggest that these two operational levels might not be separated from each other. In other words, the way our brain represents objects to support recognition might be intrinsically related to the way object representations support the different behavioral domains. Hence, state-of-the-art object recognition frameworks alone might not be sufficient to fully explain the content of our percepts and the large diversity of representations that coexist in the human brain (**Figure 4**). Evidence already present in the literature supports this shift of perspective based on the richness and often overlooked complexity of OTC representational space.

Examples that help to unveil the complexity of OTC object space and the many factors that drive this organization come from reports of representational similarities among objects that cannot be easily explained by either visual or categorical similarities. Most striking is the representational overlap for hands and tools (Bracci et al. 2012, Bracci & Peelen 2013). Hands and tools differ in many low-level visual aspects (shape, color, and motion) as well as high-level semantic properties (hands are animate, and tools are inanimate), yet their representations converge in OTC. The object recognition framework cannot explain this representational overlap, which instead can be better understood in terms of action-related relations. Namely, visual behavior that supports the way we interact with our surroundings is facilitated by proximity of representations that need constant exchange of signal information. Tools, like human hands, are action effectors that extend the functionality of our arms and are assimilated into the body schema to successfully control them (Miller et al. 2018). Amputees who regularly use prosthetic limbs to functionally overcome the missing hand recruit visual areas devoted to visual representation of the hand (van den Heiligenberg et al. 2018), although, within this region, a separated representation for the prosthesis is maintained (Maimon-Mor & Makin 2020). These results can be interpreted when considering the relevance of object recognition in the context of specific behavioral goals. In the context

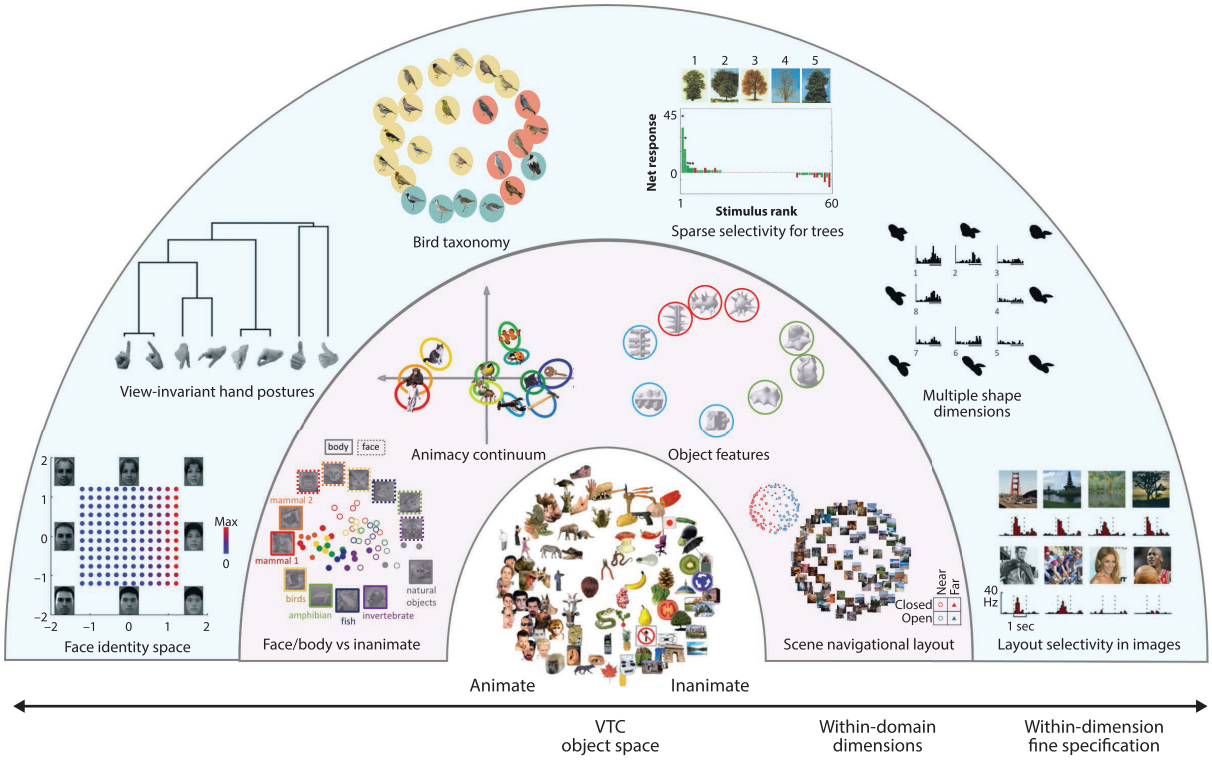

Figure 4

Illustration of findings on the representational diversity in the primate brain. The three semicircles illustrate the level of detail at which these representations have been studied. The images show spatial stimulus configurations in which stimuli that are highly similar in terms of the elicited neural response are in close proximity. (*Inner semicircle*) The structure of the large object space as obtained with stimulus-rich designs in which many dimensions covary (adapted from Kriegeskorte et al. 2008 with permission from Elsevier), which already suggests an organization in terms of animacy. (*Middle semicircle*) More detailed views of within-domain object space in studies that look into some of the important dimensions in which object space is structured. From left to right: face/body selectivity versus animacy (adapted from Ritchie et al. 2021 with permission from the author), animacy continuum (adapted from Sha et al. 2015 with permission from MIT Press), categorical shape features (object classes are clustered in terms of features such as spiky extrusions, smooth edges, or straight edges; adapted from Op de Beeck et al. 2008c), and scene spatial layout information [adapted from Kravitz et al. 2011 (CC-BY-NC-SA 4.0)]. (*Outer semicircle*) Detailed selectivity for fine within-category object distinctions. From left to right: face identity space in the monkey's face cells (adapted from Chang & Tsao 2017), viewpoint-invariant hand postures (the tree diagram built from neural similarity matrix shows that similar hand postures tend to occupy the same branch, suggesting high similarity; adapted from Bracci et al. 2018 with permission from Elsevier), bird taxonomy space in bird experts (a systematic similarity structure within the domain of birds is shown by different colors; based upon Duyck et al. 2021), sparse selectivity for tree exemplars (ranked responses of a monkey inferior temporal neuron to trees in green and nontrees in red, with the five most preferred images displayed on top; adapted from Vogels 1999 with permission from John Wiley & Sons), fine-scale shape dimensions (the response of one single inferior temporal neuron that prefers the top-left shape and whose response decreases as a function of distance in a two-dimensional shape space; adapted from Op de Beeck et al. 2001), and layout selectivity in single cells [adapted from Mormann et al. 2017 (CC BY 4.0)]. Abbreviation: VTC, ventral temporal cortex.

of object semantics, hands and tools are seen as two distal categories. On the contrary, hands and tools are typically recognized in the context of action-related computations where they share the common property of action effectors; hence this justifies their proximity in object space.

Object recognition does not happen out of context. Thus, the need to recognize an object and the need to use the content of object representations to support specific computations might

converge to an object space that maximizes distance/vicinity between representations not only in terms of recognition needs but also in terms of needs to support output behavior. This observation is also in line with a recent proposal that suggests that the ventral pathway, traditionally viewed as the object recognition pathway, might instead be organized in two separated pathways: a ventral one for object recognition and a lateral one for action recognition (Wurm & Caramazza 2022). In a similar fashion, Pitcher & Ungerleider (2021) suggest a third pathway for social-related perception. Together, these multiple lines of evidence suggest that trying to interpret the object space in the ventral pathway just in terms of diagnostic features necessary to support core object recognition goals will fail to capture the full complexity of its representational content.

An organizational space that reflects the way representations are used to support behavior is also observed in other cortices other than visual cortex. In a series of elegant studies, Graziano & Aflalo (2007) revealed how the multiple and overlapping body maps observed in motor cortex can be better understood in terms of the behavioral repertoire of the animal. That is, the topographical organization of body parts in the wider motor cortex reflects the need for the motor system to converge toward the animal's complex actions, such as hand-to-mouth or climbing/leaping movements, as opposed to the traditional view of an organization that reflects muscle proximity.

Apart from the object and action domains, other broad domains can be identified, such as the social domain and spatial/navigational domain. The social domain is possibly one of the most important cognitive domains that defines our species. Visual processing of other individuals (and animals) is fundamental to support our social skills, allowing us to understand emotions, intentions, and even the trustworthiness of our conspecifics. In OTC, representations for living entities are encoded in the lateral FG (Kanwisher et al. 1997), whereas the medial portion of FG represents objects and scenes (Epstein & Kanwisher 1998, Chao et al. 1999). This division naturally makes sense within the context of object recognition computations: It clearly separates representations allowing fast and efficient visual categorization for exemplars within categories as opposed to between categories (Grill-Spector & Weiner 2014), and it predicts the straightforward object space division between animate and inanimate entities that is often reported. However, also in this case, examples of representational similarity that are difficult to explain within the general object recognition framework are observed.

Connolly and colleagues (2012) showed that the classic animacy division might be better interpreted in terms of a continuum that spans from the most animate animals (primates) to the least animate animals (insects), where the latter are represented as closer to inanimate objects (tools) than to other animate entities (Sha et al. 2015). In a similar fashion, but in the opposite direction, it was reported that inanimate objects that share animate-like features (e.g., a coffee mug shaped as a cow) are represented as closer to real animals (e.g., cow) than to their matched inanimate objects (e.g., a coffee mug) (Bracci et al. 2019). This result is particularly striking because from the point of view of object recognition, a coffee mug with the shape of a cow is identical in all respects (object identity in primis, but also visual aspects such as shape and size) to a plain coffee mug. On the contrary, DCNNs trained on object recognition represent two mugs as being similar regardless of whether one is shaped like a cow (Bracci et al. 2019). Similar findings have been reported in the face domain. Responses in human cortex to object images that induce face illusory effects (pareidolia) confirm that the initial response in OTC to illusory face objects is strongly face-like (Wardle et al. 2020). These results show that up to a certain stage of visual processing, the relevance of diagnostic features to detect faces or features that characterize animate entities overrules the relevance of diagnostic features necessary to recognize the identity of an object.

The diagnostic features that characterize a living entity might be visual, such as the presence of eyes or a mouth, in line with recent studies showing that the presence of typical features of faces and bodies explains the animacy organization over and above the animacy continuum (Ritchie et al.

2021, Proklova & Goodale 2022). The underlying representation has to be sufficiently schematic and abstract, though, in order to capture the many different forms that such features can take, in particular when other dimensions also vary (see, e.g., Bracci & Op de Beeck 2016). Diagnostic features to detect animacy might also be inferred by goal-directed actions in the absence of animate-like visual stimuli, as shown in studies employing visual displays of geometric shapes that move in an animate fashion (Martin & Weisberg 2003, Gobbini et al. 2007). These studies show consistent activation in the lateral FG when visual displays of moving geometric shapes conveying meaningful social interactions (e.g., chasing one another) are contrasted with matched displays where the dots move in a mechanic/random way. Sensitivity to animacy-like visual and motion properties such as face-like patterns or self-initiated movements of inanimate shapes appears to be present already in human infants (Di Giorgio et al. 2017, Buiatti et al. 2019).

For humans, the necessity to recognize a living animal is fundamental for social interactions and requires the visual processing of faces and bodies and the analysis of social cues that convey identity, emotions, and posture. Since identification of someone's identity or emotion requires the analysis of the whole body, it has been proposed that face- and body-selective areas might be a rather unified node that supports person perceptual analysis (Taubert et al. 2022). In addition, processing of body part movements allows interpretation of other individuals' complex behaviors. Indeed, in contrast with interactions with inanimate objects, social interactions are dynamic and necessitate the anticipation of possible reactions of other individuals to our actions (e.g., shaking hands, hugging). The lateral OTC appears to play a special role in representing dynamic information on agents, with the posterior portion of superior temporal sulcus (STS) encoding biological motion of faces and bodies (Beauchamp et al. 2002, Grossmann & Blake 2002) as well as playing an important role in processes that support our ability to infer other people's mental states (Saxe & Wexler 2005). These is also evidence showing that the object space in OTC reflects social agency (Sha et al. 2015, Papeo et al. 2017, Thorat et al. 2019, Jozwik et al. 2021) and that the representational space for body parts reflects action-related properties (Bracci et al. 2015, 2018) as opposed to other visual dimensions. Taken as a whole, this evidence points toward a complex scenario. Clearly, attempts to identify a few ideal dimensions that aspire to capture the whole complexity of OTC object space are unrealistic. Most likely, a combination of bottom-up factors (Levy et al. 2001) and top-down factors (e.g., connectivity constrains; Mahon & Caramazza 2011, Saygin et al. 2012, Peelen & Downing 2017, Op de Beeck et al. 2019) contributes to shaping OTC representational content, in a way that allows fast readouts of object identity as well as diagnostic features relevant to support one of many possible behavioral goals.

Here, we focused upon a relatively small set of dimensions and domains that are particularly prominent in the representational object space at both the psychological and the neural level, such as animacy and faces. However, the same reasoning that behavioral goals are crucial to recognize objects can be extended toward a wider range of object categories. For instance, recent evidence shows that place-related processing is tied to the behavioral possibilities offered by a scene, with separated representations depending on whether a scene conveys action-related information such as close-up reachable spaces (e.g., sink with plates and cutlery) or spatial navigation information for far-off sceneries (Josephs & Konkle 2020). In a similar fashion, the function of an object is an important determinant for categorization (Booth & Waxman 2002, Oakes & Madole 2008). To take an example related to the affordance for actions as was already studied in the ecological approach to vision (Gibson 1979), something is a chair only if one can sit on it (Grabner et al. 2011). The behavioral goals that a category relates to not only are relevant for the action itself, a goal that is typically associated with the dorsal visual stream (Goodale & Milner 1992), but also determine the core features determining the perception, categorization, and indeed also recognition of an object.

The view proposed here—that in order to better appreciate OTC object space that supports object recognition, we need to shift our perspective to integrate the operational level of output behavior—does not mean to diminish the fundamental role that OTC plays in object recognition. On the contrary, what we suggest is that representations that sustain object recognition goals might be entangled with higher-level goal-directed representations. When processing our visual surrounding, we do so with a goal in mind, whether we are walking toward a specific destination or are looking for a friend with whom we have an appointment. In such a circumstance, the same visual scenario needs to be processed depending on the given goal. Of course, before we can walk toward and wave to our friend, we need to recognize their identity, and before we can successfully reach our destination, we need to recognize the location of our appointment relative to a mental map of the surrounding space. What we propose is that the same areas involved in exploiting this information to support behavior are also engaged in recognition, given that an object becomes relevant only in the context of a specific behavior. As William James (1890, p. 333) said, "There is no property absolutely essential to any one thing. The same property which figures as the essence of a thing on one occasion becomes a very inessential feature upon another." Furthermore, the same object might serve different computations, thus predicting differential relevance of different object features. DCNNs provide a promising avenue to better understand the role of behavioral output constraints in shaping OTC object space. In the next section we discuss current developments on this front.

5. DEEP CONVOLUTIONAL NEURAL NETWORKS NEED MORE REPRESENTATIONAL DIVERSITY

The richness of the representational space that follows from the many behavioral goals served by human object vision is a major challenge for computational efforts to model human vision. Upon closer scrutiny, the correspondences that have been found between human vision and the representations in trained DCNNs turn out to be superficial and very incomplete. DCNNs trained in general image classification develop representations that are often organized according to some of the same dimensions as human vision, such as animacy, visual shape, and aspect ratio (Bao et al. 2020, Zeman et al. 2020). However, there are a lot of domains in which these networks lack any understanding. For example, humans know thousands of object categories and thousands of individual faces, while DCNNs are typically not combining the two levels of expertise (Dobs et al. 2022). It has proven useful to study networks trained on specific goals within a particular domain of stimuli, and often such networks have been shown to be superior in terms of how they correspond to coding properties in specific domains and brain areas. Consider faces: Faces are important from a social perspective, especially certain aspects of faces such as identity, facial expression, or direction of gaze. DCNNs trained in face recognition show a good fit with neuronal responses to faces in human intracranial recordings (Grossman et al. 2019). Similarly, in the domain of scene perception, DCNNs can be trained on many specific aspects of scenes that are important for scene cognition, such as scene categorization, 3D elements, and affordance. Some of these models are particularly good at predicting representational similarity in specific scene-selective regions (Bonner & Epstein 2018, Dwivedi et al. 2021).

However, for all training regimes, and in particular for the most specific ones, we can question the level at which these networks understand the structure and content of images. The typical DCNNs trained on image classification might give the false impression that they have a human-like understanding of the content of visual images, but they fall short of that. They do not even understand what an object is in its most fundamental sense (defined as a material thing that can be seen and touched). Even for the networks that received a broad training with one thousand object categories, images are simply statistical patterns, and so might be the categories that they

have learned to differentiate. A first sign of trouble can be found in the characteristics of many adversarial images. For example, adversarial images often contain some features of an object class but no object at all (Nguyen et al. 2015). A pattern of alternating black and yellow rectangles might be classified with high confidence as an (American) school bus, even though no object is present. Often such errors are illustrative of the general tendency of DCNNs to rely more upon texture and ignore global shape, in contrast to human observers (Baker et al. 2020).

A more recent illustration was obtained through testing DCNNs and the human visual system with images that contain strong object-to-background correspondence (Bracci et al. 2022). In ImageNet as well as in real life, some objects are typically depicted against specific backgrounds. A ladybug is typically found on green leaves; yet, the leaves are not part of the ladybug. Human object recognition can be helped by using such correspondences, as shown by better and faster recognition when an object is shown in its typical context (Biederman 1972, Kaiser et al. 2019), yet objects and backgrounds are represented in different parts of the visual system. No region in the human visual pathway responds similarly to a ladybug and a leaf, not even when participants are searching for such correspondences (Bracci et al. 2022). This indicates that the coding of such correspondences is built on top of a more primary understanding that ladybugs and leaves are fundamentally different, even though they often go together. DCNN models build up a very different type of representation. The fully connected DCNN layers that in other experiments show a good correspondence with human visual cortex are not able to separate objects from backgrounds: These layers show a similar response pattern to ladybugs and to leaves. Such problematic behavior is not simply solved by extending network architecture to higher depth or recurrence. The strong mixing of representing objects and background is also seen in deep ResNet architectures, which are shown to also mimic the possibility of recurrent processing to deal with object/background clutter (Seijdel et al. 2021, Bracci et al. 2022).

There are various related approaches that focus upon tasks and signals that bring us closer to the aforementioned definition of "object." It is possible to perform the segmentation prior to feeding images into a recognition network or to enrich a DCNN architecture with segmentation routines (e.g., Zhang et al. 2020). Recent work also showed that the representational space in modern generative networks contains latent dimensions that differentially tag foreground and background (Voynov & Babenko 2020). These findings suggest that DCNNs can come up with at least some sub-solutions needed to understand the concept of object, and a more diverse training and task regime might prove relevant. The point is that networks are needed that incorporate training for multiple goals.

If even DCNNs trained on one thousand object categories struggle to account for very basic properties of images such as the presence of objects, how do we evaluate DCNNs trained in more specific domains? We can indeed question the image understanding of networks trained on specific tasks such as face identification, facial expression recognition, scene affordance, and the like. A network trained in face identification does not even know what a face is. It will respond as strongly to nonface images as it does to face images, even by definition (the most common output transformation in DCNNs is a softmax operation, which normalizes all output to 1). In human vision, recognizing a face builds on top of detecting a face and segmenting it from the background. There is a multitude of processes and representations involved when humans recognize a face, processes that are often not specific to a particular domain but that strongly influence the understanding of this domain. It is interesting that recent attempts to model the face preference in human face-selective cortex found that starting from broadly trained models (e.g., trained on one thousand ImageNet categories) perform better than domain-specific models trained, for example, in face identification (Ratan Murty et al. 2021). Apparently, a broad experience helps to simulate the general face preference of the region as a whole. However, this training regime would be

counterproductive to building up a representation that is useful for the more specific tasks for which we use the face network, such as face identification. Human face-selective cortex does it all: It can detect faces in general and supports face individuation. In conclusion, what we need to simulate the processing in domain-specific cortex such as face regions are networks trained not only in the differentiation among face images in terms of, for example, identity and emotion, but also in the detection and differentiation of faces from other categories and in other more generic processes such as object/background segmentation. To go one step further, we need to train these networks to solve real-world person-related tasks (e.g., what are these two people talking about? Do they know each other? Is the older one helping the younger one?) used in multimodal networks that integrate vision and other modalities such as language (Bernardi & Pezzelle 2021).

Thus, we lack computational systems containing representations of many different aspects of images that are developed having multiple goals in mind and within the context of a system trained for the bigger picture, where domains interact. This is only the start. If an image is already worth one thousand representations, or whatever the exact number, then what to think about movies and about multimodal input? We are certainly narrowminded when we focus here upon the perception of still images. A true understanding of images might be unlikely or even impossible without additional sources of information. One step further is image motion. Image motion is a powerful cue for image structure in human perception, such as in the Gestalt principle of common fate, and its temporal development is a key property to allow DCNNs such as predictive coding networks (PredNet; Lotter et al. 2020) to implicitly learn object and scene structure. Furthermore, multisensory convergence might be needed to come to a real understanding of "a material thing that can be seen and touched." Without this additional context a DCNN (or a human, for that matter) might never truly transcend its understanding beyond the level of learning image statistics.

6. CONCLUSION

As our brain evolved to support adaptive behavior, our visual system, too, might be at the service of our behavioral needs. Object recognition does not happen in a vacuum. Therefore, the way our visual system represents objects might be intrinsically related to the representations employed to support behavior. We suggest that to crack the code of object vision, object representations need to be investigated at multiple levels of detail, and their multidimensional nature should be taken seriously. Furthermore, we need to consider the interaction between different domains: Even a domain-specific system like the face recognition system has evolved in the context of a much larger visual system and interfaces with many domains at multiple levels. Likewise, a computational model aimed at fully capturing human cognition in a particular domain will also need to incorporate how this domain interacts with other domains. It is the rich and across-domain representational zoo that grants the human brain its exceptional powers in multipurpose information processing.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We would like to thank Scott Fairhall and Brendan Ritchie for their useful feedback on the manuscript. H.P.O. is supported by FWO/FNRS EOS (Excellence of Science) grant HumVisCat (No. 30991544), FWO research project G0D3322N, and KU Leuven project IDN/21/010.

LITERATURE CITED

Arcaro MJ, Livingstone MS. 2017. A hierarchical, retinotopic proto-organization of the primate visual system at birth. *eLife* 6:e26196

Arcaro MJ, Livingstone MS. 2021. On the relationship between maps and domains in inferotemporal cortex. *Nat. Rev. Neurosci.* 22(9):573–83

Attneave F. 1957. Physical determinants of the judged complexity of shapes. *J. Exp. Psychol.* 53(4):221–27

Avberšek LK, Zeman A, Op de Beeck HP. 2021. Training for object recognition with increasing spatial frequency: a comparison of deep learning with human vision. *J. Vis.* 21(10):14

Baker N, Lu H, Erlikhman G, Kellman PJ. 2020. Local features and global shape information in object classification by deep convolutional neural networks. *Vis. Res.* 172:46–61

Baldassi C, Alemi-Neissi A, Pagan M, DiCarlo JJ, Zecchina R, Zoccolan D. 2013. Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons. *PLOS Comput. Biol.* 9(8):e1003167

Bao P, She L, McGill M, Tsao DY. 2020. A map of object space in primate inferotemporal cortex. *Nature* 583(7814):103–8

Barton JJ, Press DZ, Keenan JP, O'Connor M. 2002. Lesions of the fusiform face area impair perception of facial configuration in prosopagnosia. *Neurology* 58(1):71–78

Baylis GC, Rolls ET, Leonard CM. 1987. Functional subdivisions of the temporal lobe neocortex. *J. Neurosci.* 7(2):330–42

Beauchamp MS, Lee KE, Haxby JV, Martin A. 2002. Parallel visual motion processing streams for manipulable objects and human movements. *Neuron* 34(1):149–59

Bernardi R, Pezzelle S. 2021. Linguistic issues behind visual question answering. *Lang. Linguist. Compass* 15(6):e12417

Biederman I. 1972. Perceiving real-world scenes. *Science* 177(4043):77–80

Biederman I. 1987. Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94(2):115–47

Blauch NM, Behrmann M, Plaut DC. 2022. A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *PNAS* 119(3):e2112566119

Bonner MF, Epstein RA. 2018. Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLOS Comput. Biol.* 14(4):e1006111

Booth AE, Waxman S. 2002. Object names and object functions serve as cues to categories for infants. *Dev. Psychol.* 38(6):948–57

Bracci S, Caramazza A, Peelen MV. 2015. Representational similarity of body parts in human occipitotemporal cortex. *J. Neurosci.* 35(38):12977–85

Bracci S, Caramazza A, Peelen MV. 2018. View-invariant representation of hand postures in the human lateral occipitotemporal cortex. *Neuroimage* 181:446–52

Bracci S, Cavina-Pratesi C, Ietswaart M, Caramazza A, Peelen MV. 2012. Closely overlapping responses to tools and hands in left lateral occipitotemporal cortex. *J. Neurophysiol.* 107(5):1443–56

Bracci S, Ietswaart M, Peelen MV, Cavina-Pratesi C. 2010. Dissociable neural responses to hands and non-hand body parts in human left extrastriate visual cortex. *J. Neurophysiol.* 103(6):3389–97

Bracci S, Mraz J, Zeman A, Leys G, Op de Beeck HP. 2022. The representational hierarchy in human and artificial visual systems in the presence of object-scene regularities. bioRxiv 456197. **https://doi.org/10.1101/2021.08.13.456197**

Bracci S, Op de Beeck HP. 2016. Dissociations and associations between shape and category representations in the two visual pathways. *J. Neurosci.* 36(2):432–44

Bracci S, Peelen MV. 2013. Body and object effectors: the organization of object representations in high-level visual cortex reflects body–object interactions. *J. Neurosci.* 33(46):18247–58

Bracci S, Ritchie JB, Kalfas I, Op de Beeck HP. 2019. The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. *J. Neurosci.* 39(33):6513–25

Bracci S, Ritchie JB, Op de Beeck HP. 2017. On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia* 105:153–64

Buiatti M, Di Giorgio E, Piazza M, Polloni C, Menna G, et al. 2019. Cortical route for facelike pattern processing in human newborns. *PNAS* 116(10):4625–30

Canário N, Jorge L, Silva ML, Soares MA, Castelo-Branco M. 2016. Distinct preference for spatial frequency content in ventral stream regions underlying the recognition of scenes, faces, bodies and other objects. *Neuropsychologia* 87:110–19

Caramazza A, Shelton JR. 1998. Domain-specific knowledge systems in the brain: the animate-inanimate distinction. *J. Cogn. Neurosci.* 10(1):1–34

Chang L, Tsao DY. 2017. The code for facial identity in the primate brain. *Cell* 169(6):1013–28

Chao LL, Haxby JV, Martin A. 1999. Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nat. Neurosci.* 2(10):913–19

Chklovskii DB, Koulakov AA. 2004. Maps in the brain: What can we learn from them? *Annu. Rev. Neurosci.* 27:369–92

Cohen L, Lehéricy S, Chochon F, Lemer C, Rivaud S, Dehaene S. 2002. Language-specific tuning of visual cortex? Functional properties of the Visual Word Form Area. *Brain* 125(5):1054–69

Connolly AC, Guntupalli JS, Gors J, Hanke M, Halchenko YO, et al. 2012. The representation of biological classes in the human brain. *J. Neurosci.* 32(8):2608–18

Cox DD, Savoy RL. 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* 19(2):261–70

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. 2009. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–55. New York: IEEE

Desimone R, Albright TD, Gross CG, Bruce C. 1984. Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.* 4(8):2051–62

Di Giorgio E, Lunghi M, Simion F, Vallortigara G. 2017. Visual cues of motion that trigger animacy perception at birth: the case of self-propulsion. *Dev. Sci.* 20(4):e12394

DiCarlo JJ, Maunsell JH. 2003. Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *J. Neurophysiol.* 89(6):3264–78

DiCarlo JJ, Zoccolan D, Rust NC. 2012. How does the brain solve visual object recognition? *Neuron* 73(3):415–34

Dobs K, Martinez J, Kell AJ, Kanwisher N. 2022. Brain-like functional specialization emerges spontaneously in deep neural networks. *Sci. Adv.* 8(11):eabl8913

Downing PE, Jiang Y, Shuman M, Kanwisher N. 2001. A cortical area selective for visual processing of the human body. *Science* 293(5539):2470–73

Dujmović M, Malhotra G, Bowers JS. 2020. What do adversarial images tell us about human vision? *eLife* 9:e55978

Duyck S, Martens F, Chen CY, Op de Beeck HP. 2021. How visual expertise changes representational geometry: a behavioral and neural perspective. *J. Cogn. Neurosci.* 33(12):2461–76

Dwivedi K, Bonner MF, Cichy RM, Roig G. 2021. Unveiling functions of the visual cortex using task-specific deep neural networks. *PLOS Comput. Biol.* 17(8):e1009267

Elmoznino E, Bonner MF. 2022. High-performing neural network models of visual cortex benefit from high latent dimensionality. bioRxiv 499969. **https://doi.org/10.1101/2022.07.13.499969**

Epstein R, Kanwisher N. 1998. A cortical representation of the local visual environment. *Nature* 392(6676):598–601

Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv:1811.12231 [cs.CV]

Gibson JJ. 1979. *The Ecological Approach to Visual Perception*. New York: Psychol. Press

Gobbini MI, Koralek AC, Bryan RE, Montgomery KJ, Haxby JV. 2007. Two takes on the social brain: a comparison of theory of mind tasks. *J. Cogn. Neurosci.* 19(11):1803–14

Goffaux V, Dakin S. 2010. Horizontal information drives the behavioral signatures of face processing. *Front. Psychol.* 1:143

Gomez J, Natu V, Jeska B, Barnett M, Grill-Spector K. 2018. Development differentially sculpts receptive fields across early and high-level human visual cortex. *Nat. Commun.* 9:788

Goodale MA, Milner AD. 1992. Separate visual pathways for perception and action. *Trends Neurosci.* 15(1):20–25

Grabner H, Gall J, Van Gool L. 2011. What makes a chair a chair? In *CVPR 2011: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1529–36. New York: IEEE

Graziano MS, Aflalo TN. 2007. Mapping behavioral repertoire onto the cortex. *Neuron* 56(2):239–51

Grill-Spector K, Kushnir T, Edelman S, Avidan G, Itzchak Y, Malach R. 1999. Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron* 24(1):187–203

Grill-Spector K, Kushnir T, Hendler T, Edelman S, Itzchak Y, Malach R. 1998. A sequence of object-processing stages revealed by fMRI in the human occipital lobe. *Hum. Brain Mapp.* 6(4):316–28

Grill-Spector K, Weiner KS. 2014. The functional architecture of the ventral temporal cortex and its role in categorization. *Nat. Rev. Neurosci.* 15(8):536–48

Gross CG, Rocha-Miranda CD, Bender DB. 1972. Visual properties of neurons in inferotemporal cortex of the macaque. *J. Neurophysiol.* 35(1):96–111

Grossman ED, Blake R. 2002. Brain areas active during visual perception of biological motion. *Neuron* 35(6):1167–75

Grossman S, Gaziv G, Yeagle EM, Harel M, Mégevand P, et al. 2019. Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nat. Commun.* 10:4934

Güçlü U, van Gerven MA. 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* 35(27):10005–14

Hasson U, Levy I, Behrmann M, Hendler T, Malach R. 2002. Eccentricity bias as an organizing principle for human high-order object areas. *Neuron* 34(3):479–90

Hebart MN, Zheng CY, Pereira F, Baker CI. 2020. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nat. Hum. Behav.* 4(11):1173–85

Hoffman DD, Richards WA. 1984. Parts of recognition. *Cognition* 18(1–3):65–96

Hong H, Yamins DL, Majaj NJ, DiCarlo JJ. 2016. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* 19(4):613–22

Hu JM, Song XM, Wang Q, Roe AW. 2020. Curvature domains in V4 of macaque monkey. *eLife* 9:e57261

Hubel DH, Wiesel TN. 1968. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195(1):215–43

Hung CP, Kreiman G, Poggio T, DiCarlo JJ. 2005. Fast readout of object identity from macaque inferior temporal cortex. *Science* 310(5749):863–66

Jagadeesh AV, Gardner JL. 2022. Texture-like representation of objects in human visual cortex. *PNAS* 119(17):e2115302119

Jain N, Wang A, Henderson MM, Lin R, Prince JS, et al. 2022. Food for thought: selectivity for food in human ventral visual cortex. bioRxiv 492983. **https://doi.org/10.1101/2022.05.22.492983**

James W. 1890. *The Principles of Psychology*, Vol. 1. London: Macmillan

Jenkins R, Dowsett AJ, Burton AM. 2018. How many faces do people know? *Proc. R. Soc. B* 285(1888):20181319

Josephs EL, Konkle T. 2020. Large-scale dissociations between views of objects, scenes, and reachable-scale environments in visual cortex. *PNAS* 117(47):29354–62

Jozwik KM, Najarro E, van den Bosch JJ, Charest I, Kriegeskorte N, Cichy RM. 2021. Disentangling five dimensions of animacy in human brain and behaviour. bioRxiv 459854. **https://doi.org/10.1101/2021.09.12.459854**

Kaiser D, Quek GL, Cichy RM, Peelen MV. 2019. Object vision in a structured world. *Trends Cogn. Sci.* 23(8):672–85

Kanwisher N, McDermott J, Chun MM. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* 17(11):4302–11

Kar K, Kubilius J, Schmidt K, Issa EB, DiCarlo JJ. 2019. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* 22(6):974–83

Kayaert G, Biederman I, Op de Beeck HP, Vogels R. 2005. Tuning for shape dimensions in macaque inferior temporal cortex. *Eur. J. Neurosci.* 22(1):212–24

Keller TA, Gao Q, Welling M. 2021. Modeling category-selective cortical regions with topographic variational autoencoders. arXiv:2110.13911 [q-bio.NC]

Khosla M, Murty NAR, Kanwisher NG. 2022. A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. bioRxiv 496922. **https://doi.org/10.1101/2022.06.21.496922**

Kietzmann TC, Spoerer CJ, Sörensen LK, Cichy RM, Hauk O, Kriegeskorte N. 2019. Recurrence is required to capture the representational dynamics of the human visual system. *PNAS* 116(43):21854–63

Kim E, Rego J, Watkins Y, Kenyon GT. 2020. Modeling biological immunity to adversarial examples. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4666–75. New York: IEEE

Konkle T, Alvarez GA. 2022. Beyond category-supervision: computational support for domain-general pressures guiding human visual system representation. *Nat. Commun.* 13:491

Konkle T, Oliva A. 2012. A real-world size organization of object responses in occipitotemporal cortex. *Neuron* 74(6):1114–24

Kourtzi Z, Kanwisher N. 2001. Representation of perceived object shape by the human lateral occipital complex. *Science* 293(5534):1506–9

Kravitz DJ, Peng CS, Baker CI. 2011. Real-world scene representations in high-level visual cortex: It's the spaces more than the places. *J. Neurosci.* 31(20):7322–33

Kravitz DJ, Saleem KS, Baker CI, Ungerleider LG, Mishkin M. 2013. The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends Cogn. Sci.* 17(1):26–49

Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, et al. 2008. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60(6):1126–41

Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inform. Proc. Syst.* 25:1097–105

Kubilius J, Bracci S, Op de Beeck HP. 2016. Deep neural networks as a computational model for human shape sensitivity. *PLOS Comput. Biol.* 12(4):e1004896

Ito M, Tamura H, Fujita I, Tanaka K. 1995. Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol.* 73(1):218–26

Laiacona M, Barbarotto R, Capitani E. 1993. Perceptual and associative knowledge in category specific impairment of semantic memory: a study of two cases. *Cortex* 29(4):727–40

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521(7553):436–44

Lee H, Margalit E, Jozwik KM, Cohen MA, Kanwisher N, et al. 2020. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. bioRxiv 185116. **https://doi.org/10.1101/2020.07.09.185116**

Levy I, Hasson U, Avidan G, Hendler T, Malach R. 2001. Center–periphery organization of human object areas. *Nat. Neurosci.* 4(5):533–39

Li SPD, Bonner MF. 2021. Tuning in scene-preferring cortex for mid-level visual features gives rise to selectivity across multiple levels of stimulus complexity. bioRxiv 461733. **https://doi.org/10.1101/2021.09.24.461733**

Lindsay GW. 2021. Convolutional neural networks as a model of the visual system: past, present, and future. *J. Cogn. Neurosci.* 33(10):2017–31

Long B, Yu CP, Konkle T. 2018. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *PNAS* 115(38):E9015–24

Lotter W, Kreiman G, Cox D. 2020. A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nat. Mach. Intell.* 2(4):210–19

Mahon BZ, Caramazza A. 2011. What drives the organization of object knowledge in the brain? *Trends Cogn. Sci.* 15(3):97–103

Maimon-Mor RO, Makin TR. 2020. Is an artificial limb embodied as a hand? Brain decoding in prosthetic limb users. *PLOS Biol.* 18(6):e3000729

Malach R, Levy I, Hasson U. 2002. The topography of high-order human object areas. *Trends Cogn. Sci.* 6(4):176–84

Malach R, Reppas JB, Benson RR, Kwong KK, Jiang H, et al. 1995. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *PNAS* 92(18):8135–39

Marr D. 1980. Visual information processing: the structure and creation of visual representations. *Philos. Trans. R. Soc. B* 290(1038):199–218

Marr D, Nishihara HK. 1978. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. B* 200(1140):269–94
Martin A, Weisberg J. 2003. Neural foundations for understanding social and mechanical concepts. *Cogn. Neuropsychol.* 20(3–6):575–87
Mehrer J, Spoerer CJ, Jones EC, Kriegeskorte N, Kietzmann TC. 2021. An ecologically motivated image dataset for deep learning yields better models of human vision. *PNAS* 118(8):e2011417118
Miller LE, Montroni L, Koun E, Salemme R, Hayward V, Farnè A. 2018. Sensing with tools extends somatosensory processing beyond the body. *Nature* 561(7722):239–42
Mishkin M, Ungerleider LG. 1982. Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behav. Brain Res.* 6(1):57–77
Morgenstern Y, Hartmann F, Schmidt F, Tiedemann H, Prokott E, et al. 2021. An image-computable model of human visual shape similarity. *PLOS Comput. Biol.* 17(6):e1008981
Mormann F, Kornblith S, Cerf M, Ison MJ, Kraskov A, et al. 2017. Scene-selective coding by single neurons in the human parahippocampal cortex. *PNAS* 114(5):1153–58
Nasr S, Echavarria CE, Tootell RB. 2014. Thinking outside the box: Rectilinear shapes selectively activate scene-selective cortex. *J. Neurosci.* 34(20):6721–35
Nguyen A, Yosinski J, Clune J. 2015. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 427–36. New York: IEEE
Oakes LM, Madole KL. 2008. Function revisited: how infants construe functional features in their representation of objects. *Adv. Child Dev. Behav.* 36:135–85
Op de Beeck HP, Deutsch JA, Vanduffel W, Kanwisher NG, DiCarlo JJ. 2008a. A stable topography of selectivity for unfamiliar shape classes in monkey inferior temporal cortex. *Cereb. Cortex* 18(7):1676–94
Op de Beeck HP, Haushofer J, Kanwisher NG. 2008b. Interpreting fMRI data: maps, modules and dimensions. *Nat. Rev. Neurosci.* 9(2):123–35
Op de Beeck HP, Pillet I, Ritchie JB. 2019. Factors determining where category-selective areas emerge in visual cortex. *Trends Cogn. Sci.* 23(9):784–97
Op de Beeck HP, Torfs K, Wagemans J. 2008c. Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *J. Neurosci.* 28(40):10111–23
Op De Beeck HP, Vogels R. 2000. Spatial sensitivity of macaque inferior temporal neurons. *J. Comp. Neurol.* 426(4):505–18
Op De Beeck HP, Wagemans J, Vogels R. 2001. Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat. Neurosci.* 4(12):1244–52
Papeo L, Stein T, Soto-Faraco S. 2017. The two-body inversion effect. *Psychol. Sci.* 28(3):369–79
Peelen MV, Downing PE. 2017. Category selectivity in human visual cortex: beyond visual object recognition. *Neuropsychologia* 105:177–83
Pitcher D, Charles L, Devlin JT, Walsh V, Duchaine B. 2009. Triple dissociation of faces, bodies, and objects in extrastriate cortex. *Curr. Biol.* 19(4):319–24
Pitcher D, Ungerleider LG. 2021. Evidence for a third visual pathway specialized for social perception. *Trends Cogn. Sci.* 25(2):100–10
Powell LJ, Kosakowski HL, Saxe R. 2018. Social origins of cortical face areas. *Trends Cogn. Sci.* 22(9):752–63
Proklova D, Goodale MA. 2022. The role of animal faces in the animate-inanimate distinction in the ventral temporal cortex. *Neuropsychologia* 169:108192
Proklova D, Kaiser D, Peelen MV. 2016. Disentangling representations of object shape and object category in human visual cortex: the animate–inanimate distinction. *J. Cogn. Neurosci.* 28(5):680–92
Rajimehr R, Devaney KJ, Bilenko NY, Young JC, Tootell RB. 2011. The "parahippocampal place area" responds preferentially to high spatial frequencies in humans and monkeys. *PLOS Biol.* 9(4):e1000608
Ratan Murty NA, Bashivan P, Abate A, DiCarlo JJ, Kanwisher N. 2021. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nat. Commun.* 12:5540
Riesenhuber M, Poggio T. 1999. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2(11):1019–25
Ritchie JB, Zeman AA, Bosmans J, Sun S, Verhaegen K, Op de Beeck HP. 2021. Untangling the animacy organization of occipitotemporal cortex. *J. Neurosci.* 41(33):7103–19

Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P. 1976. Basic objects in natural categories. *Cogn. Psychol.* 8(3):382–439

Rosenke M, van Hoof R, van den Hurk J, Grill-Spector K, Goebel R. 2021. A probabilistic functional atlas of human occipito-temporal visual cortex. *Cereb. Cortex* 31(1):603–19

Saxe R, Wexler A. 2005. Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia* 43(10):1391–99

Saygin ZM, Osher DE, Koldewyn K, Reynolds G, Gabrieli JD, Saxe RR. 2012. Anatomical connectivity patterns predict face selectivity in the fusiform gyrus. *Nat. Neurosci.* 15(2):321–27

Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, et al. 2020. Brain-score: Which artificial neural network for object recognition is most brain-like? bioRxiv 407007. **https://doi.org/10.1101/407007**

Seijdel N, Loke J, Van de Klundert R, Van der Meer M, Quispel E, et al. 2021. On the necessity of recurrent processing during object recognition: It depends on the need for scene segmentation. *J. Neurosci.* 41(29):6281–89

Sha L, Haxby JV, Abdi H, Guntupalli JS, Oosterhof NN, et al. 2015. The animacy continuum in the human ventral vision pathway. *J. Cogn. Neurosci.* 27(4):665–78

Shepard RN, Chipman S. 1970. Second-order isomorphism of internal representations: shapes of states. *Cogn. Psychol.* 1(1):1–17

Taubert J, Ritchie JB, Ungerleider LG, Baker CI. 2022. One object, two networks? Assessing the relationship between the face and body-selective regions in the primate visual system. *Brain Struct. Funct.* 227(4):1423–38

Thorat S, Proklova D, Peelen MV. 2019. The nature of the animacy organization in human ventral temporal cortex. *eLife* 8:e47142

Van den Heiligenberg FM, Orlov T, Macdonald SN, Duff EP, Henderson Slater D, et al. 2018. Artificial limb representation in amputees. *Brain* 141(5):1422–33

Vogels R. 1999. Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study. *Eur. J. Neurosci.* 11(4):1239–55

Voynov A, Babenko A. 2020. Unsupervised discovery of interpretable directions in the GAN latent space. *PMLR* 119:9786–96

Yovel G, Kanwisher N. 2005. The neural basis of the behavioral face-inversion effect. *Curr. Biol.* 15(24):2256–62

Wardle SG, Taubert J, Teichmann L, Baker CI. 2020. Rapid and dynamic processing of face pareidolia in the human brain. *Nat. Commun.* 11:4518

Wurm MF, Caramazza A. 2022. Two "what" pathways for action and object recognition. *Trends Cogn. Sci.* 26(2):103–16

Zeman AA, Ritchie JB, Bracci S, de Beeck HO. 2020. Orthogonal representations of object shape and category in deep convolutional neural networks and human visual cortex. *Sci. Rep.* 10(1):2453

Zhang M, Tseng C, Kreiman G. 2020. Putting visual object recognition in context. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12985–94. New York: IEEE