Annual Review of Vision Science

How Shape Perception Works, in Two Dimensions and Three Dimensions

Kristina J. Nielsen and Charles E. Connor

Krieger Mind/Brain Institute and Department of Neuroscience, Johns Hopkins University, Baltimore, Maryland, USA; email: Knielse4@jhu.edu, connor@jhu.edu

Annu. Rev. Vis. Sci. 2024. 10:47–68

First published as a Review in Advance on June 7, 2024

The *Annual Review of Vision Science* is online at vision.annualreviews.org

https://doi.org/10.1146/annurev-vision-112823-031607

Keywords

visual cortex, ventral pathway, object recognition, shape, neural coding, primate

Abstract

The ventral visual pathway transforms retinal images into neural representations that support object understanding, including exquisite appreciation of precise 2D pattern shape and 3D volumetric shape. We articulate a framework for understanding the goals of this transformation and how they are achieved by neural coding at successive ventral pathway stages. The critical goals are (*a*) radical compression to make shape information communicable across axonal bundles and storable in memory, (*b*) explicit coding to make shape information easily readable by the rest of the brain and thus accessible for cognition and behavioral control, and (*c*) representational stability to maintain consistent perception across highly variable viewing conditions. We describe how each transformational step in ventral pathway vision serves one or more of these goals. This three-goal framework unifies discoveries about ventral shape processing into a neural explanation for our remarkable experience of shape as a vivid, richly detailed aspect of the natural world.

INTRODUCTION

Vision is one of the great computational feats of the brain. One of the most important aspects of visual information is shape, of both 2D patterns on surfaces and 3D volumes of objects. Shape perception is a basis for object recognition, but it encompasses far richer experiential understanding and far more information than the few bits in a categorical object label. In fact, shape perception is just as vivid and informative for things that we have never seen before, like abstract sculptures, for which we often have no labels. Conversely, recognition occurs for faces without detailed awareness of the highly complex, minute structural differences on which it is based (Hesse & Tsao 2020).

A substantial portion of neural information processing resources in the ventral visual pathway is devoted to deriving shape and other object information from retinal images. This elucidation of real structures in the physical world, from the confusing torrent of retinal information that they produce, seems so effortless that we take it for granted. However, the underlying neural processing is massively complex, only dimly understood, and yet to be duplicated by artificial vision. Artificial networks can label object categories in photographs (Krizhevsky et al. 2012, Zhuang et al. 2021), but labels are just a tiny add-on to our deep, detailed visual knowledge of the real, physical world.

Shape processing must accomplish at least three goals, in parallel, to transform retinal inputs into neural representations that support shape understanding. These goals are accomplished through stepwise transformations in successively more anterior stages of the ventral pathway. The first goal is radical compression of megapixel retinal inputs into efficient codes compact enough to transmit to other parts of the brain and store in memory. Attneave (1954) and Barlow (1961) were the first to recognize the importance of compressing visual information, especially by eliminating redundancies in retinal information. Subsequent discoveries have continued to bear out the compressive nature of visual processing (Carlson et al. 2011, Olshausen & Field 1996, Vinje & Gallant 2000).

By compression, we mean any transformation that ultimately reduces the total number of signals required to represent shape information, still in exquisite detail, at the final stages of the ventral pathway. Compression of shape information specifically, although not of visual information generally, must be relatively lossless. Compression can involve elimination or lower-resolution transfer of visual information that is redundant or irrelevant to shape perception. More often, though, it depends on transformation of many particulate signals related to tiny parts of the retinal image into fewer signals related to larger, geometrically definable parts of shapes that comprise those particles in a regular way (e.g., curvature). The strong regularities of 3D objects and 2D patterns in the physical world, due to physics, biology, and artifice, make this a powerful compression strategy and one that simultaneously produces more explicit signals for the real-world shapes that evoke retinal images.

Compression in our sense of the word is related but not equivalent to sparseness, which can denote either the fraction of activated neurons in a population (population sparseness) or the frequency of firing of one neuron across time (lifetime sparseness) (Carlson et al. 2011, Vinje & Gallant 2000). (More specifically, sparseness is the kurtosis of the neural activity histogram, i.e., the tendency toward few large activations and many small activations.) Greater sparseness (less neural activation) in neural signaling is regarded as a strategy for conserving metabolic resources. For example, in the retina, transformation from widespread activity across cones to more selective activation of retinal ganglion cells (see below) results in a sparser representation, i.e., a lower fraction of active retinal ganglion cells, and a lower metabolic cost per neuron. However, for our purposes, in understanding shape representation, the important advantage is that it results in a smaller absolute number of signals to process downstream in the ventral pathway and a

more compact, transmittable, storable, and readable representation, regardless of how sparseness changes. Sparseness is a fractional relationship between neural activations and population size (or time). Compression (in the sense of fewer total signals, the numerator in sparseness metrics) can coincide with either increased or decreased sparseness depending on how population size (the denominator) changes between processing stages.

The second parallel goal is transforming the confusing, highly implicit (hidden) information in retinal signals into explicit (easy to read or decode) information about real-world physical structure. As neural information becomes increasingly explicit along the ventral pathway, its relationship to visual cognition becomes closer and more causal (Afraz et al. 2006, Marois et al. 2004, Moeller et al. 2017, Sheinberg & Logothetis 1997, Tong et al. 1998, Verhoef et al. 2012). Explicit signals can be rapidly and easily decoded by other parts of the brain responsible for verbal description; physical interaction; and intuitions about object meaning, functionality, affordance, and value. Easy decoding usually requires only a simple, weighted sum across available signals, with weights varying depending on the specific information to be extracted (Hung et al. 2005; Pouget et al. 2002, 2003).

The third parallel goal is to make these compressed, explicit representations stable across the infinite variety of viewing conditions in real-world vision, which produce endlessly changing retinal images based on the same underlying patterns and objects. Stability means the existence of closely related neural representations across changes in position, size, orientation, lighting, posture, color, texture, etc. Some consistencies, particularly across orientation and posture, must be learned through visual experience (Li & DiCarlo 2008, Logothetis & Pauls 1995). By stability, we do not mean invariance in the responses of individual neurons, since these vary widely as a function of many image characteristics, e.g., contrast. We mean consistency of the shape representation across the neural population. A consistent representation is a response pattern across neurons that conveys the same shape information to downstream neurons, even if the individual signals that make up that pattern are weaker or stronger depending on contrast or some other factor. It is the response pattern, not the pattern magnitude, that conveys shape information.

These three aspects of stepwise transformation in the ventral visual pathway are responsible for our deep, immediate, and consistent appreciation of real-world shape. In this review, we describe how these goals are achieved through information transformations beginning in the retina and ending in the anterior temporal lobe cortex. The specific nature of these transformations determines the kinds of real-world information to which we have cognitive access. Visual information begins as 2D patterns encoded by the retinal photoreceptor array, but we do not have direct access to those patterns. Instead, we have access to the real-world structure of the patterns and objects that produce those images. This real-world knowledge lacks the full-field, absolute spatial precision of 2D projections onto the retina.

Instead, visual knowledge emphasizes detailed understanding of 2D and 3D object structure in terms of fragmentary geometry and spatial composition in relative reference frames (see below). This underlies our ability to describe object shape in extremely precise terms. It supports our appreciation of the smallest structural differences between objects that reflect their identities (my dog versus others of the same breed), histories (what kind of fights that dog must have gotten into), mood and intentions (whether that dog is happy, sad, angry, frightened, friendly, or ready to attack), and value (whether that dog is a Westminster champion or a rescue). We discuss how neural shape representations define and constrain this structural, geometric knowledge of the real world, beginning in the retina. Beyond the level of area V2, our review is limited to the ventral visual pathway, which is thought to be responsible for detailed shape vision and declarative knowledge about shape. This review does not cover the substantial literature on shape processing, especially

in three dimensions, by the dorsal pathway, which is thought to be more related to action guidance, e.g., object grasp, but which may also have critical interrelationships with ventral pathway shape processing (Rosenberg et al. 2023).

CENTER–SURROUND RECEPTIVE FIELDS IN THE RETINA

The optical image in the retina is transduced into an isomorphic 2D pattern of photoreceptor activations. Under most viewing conditions, this pattern is represented by cone receptors, which encode brightness, as well as spectral wavelength differences that enable tritanopic color vision, at minute locations in the optical image. The neural image copy is distributed over approximately 5 million cones in each human retina, mostly crowded into the fovea for highest-resolution central vision (Curcio & Allen 1990, Curcio et al. 1987).

The enormous signal carried by full-field cone activation patterns undergoes a major compression in the transformation from cones to retinal ganglion cells, mediated by the connecting bipolar cells and the lateral modulation of those connections by horizontal and amacrine cells (Field & Chichilnisky 2007). The result of the connecting circuitry is the antagonistic center–surround receptive field structure of retinal ganglion cells discovered by Kuffler (1952). The antagonistic brightness or color sensitivities of center and surround regions are balanced when they receive equivalent illumination from a spatially extended source. Thus, retinal ganglion cells respond almost exclusively to image regions with spatial brightness and/or color contrast, producing a major reduction in the total number of prominent signals representing the image. These are the regions that carry the most visual information, including information about shape. Regions of visual space that produce unvarying cone activations, like painted walls or clear skies, provide little high-resolution information to the downstream ventral pathway. The cone to retinal ganglion cell transformation involves no loss of spatial resolution in the fovea out to 1° eccentricity, where each cone receptive field is matched by the center fields of two midget retinal ganglion cells, one on center and one off-center. The midget retinal ganglion cell center field to cone field ratio remains above 1 out to 5° eccentricity but becomes much lower in the periphery, producing a loss of peripheral spatial resolution (Watson 2014).

ORIENTATION AND DISPARITY TUNING IN V1

Compression often depends on summarizing many spatially particulate signals at a lower level with a more efficient code for their overall pattern at a higher level. We do not live in a world of random pixel patterns, but in a world of surfaces and boundaries formed by physical, biological, and artificial processes. The smoothness of these surfaces and boundaries means that image contrast patterns are likewise smooth and extended. This smoothness usually has an orientation, corresponding to the angle of an object or pattern boundary or to the angle of isoluminant continuity along shading patterns on surfaces.

This oriented contrast is the nature of visual information in the natural world on the scale of V1 (primary visual cortex) receptive fields, which can be small fractions of a degree of visual angle at the fovea (Adams & Horton 2003, Gattass et al. 1987). This oriented contrast, carried by a continuous line of particulate signals from multiple retinal ganglion cells, is compressed in V1 into a representation of the line's orientation. As first described by Hubel & Wiesel (1962, 1968) based on experiments in cats and monkeys, V1 neurons respond to spatially elongated contrast patterns in a narrow range of orientations by virtue of their elongated, parallel receptive field regions with alternating sensitivities to bright versus dark and/or contrasting colors. These receptive fields may be constructed in part by linear summation of parallel rows of center–surround input (relayed via the lateral geniculate nucleus of the thalamus) to form linear simple cell spatial filters (Hubel

& Wiesel 1962, Reid & Alonso 1995). Complex cells have similar but nonlinear filter properties (Hubel & Wiesel 1962, 1968), responding to elongated contrast within a narrow orientation range, regardless of phase differences (spatial shifts of bright and dark patterns in the contrast direction, orthogonal to the elongation axis).

In addition to summarizing elongated contrast regions in terms of orientation, V1 neurons also summarize the orthogonal spread of contrast in terms of spatial frequency content (De Valois et al. 1982). Neurons tuned for high preferred spatial frequencies, i.e., oriented contrast variations with small cycle widths like fine wale corduroy, provide compressed information about spatially sharp contrasts, e.g., at 2D pattern boundaries or 3D object self-occlusion boundaries. Neurons tuned for low spatial frequencies, i.e., oriented contrast variations with large cycle widths like corrugated roofing, provide compressed information about gradual contrasts, e.g., of shading patterns on 3D surfaces. The mathematical effectiveness of this transformation on the scale of V1 receptive fields is evidenced by the reliable emergence of simple cell–like filters in artificial vision networks constrained to compress image information (Krizhevsky et al. 2012, Olshausen & Field 1996). In addition to compression, the V1 transformation provides a first step in position stability, through the phase invariance of complex cells, and the first explicit information about a geometric concept that we understand, orientation (tilt or angle).

V1 receptive fields are also sensitive to binocular image disparities between inputs from the two eyes (Cumming & Parker 1999, Poggio et al. 1985). Contrast patterns at the plane of fixation are identical between the two eyes, but nearer patterns have crossed image disparities, and farther patterns have uncrossed disparities. V1 neurons with slightly shifted sensitivities to the two eye images extract this highly implicit information, producing signals about position in the third dimension, depth. This is a first step toward explicit representation of the 3D world based on 2D retinal projection processing (Rosenberg et al. 2023).

EXPLICIT SIGNALS FOR IMPLICIT BOUNDARY INFORMATION IN V2

Brightness and color contrasts are not the only sources of structural information in visual images. Both 2D boundary shapes and 3D surface shapes can be implied by surrounding structures that they appear to partially occlude, producing boundary and surface percepts in image regions with no brightness or color contrast. Many neurons in monkey V2 exhibit orientation tuning for these illusory contours, implied by both surrounding overlapped structures and texture borders (Peterhans & von der Heydt 1989, von der Heydt & Peterhans 1989, von der Heydt et al. 1984). We are also sensitive to implicit boundaries between texture regions based solely on differential binocular disparities of their texture elements. This difference is perceived as a depth change at the boundary. V2 neurons also exhibit tuning for orientation of these implicit, disparity-based depth boundaries (von der Heydt et al. 2000).

For borders produced by normal depth differences between objects and surfaces, we can perceive which surface is in front and which is in back based on a variety of overlap cues. In connection with this ability, many V2 neurons signal border ownership, responding differentially to borders with the foreground object on one side of the boundary versus the other (Zhou et al. 2000). In this case and those mentioned above, V2 represents explicit boundary information based on implicit image cues.

DERIVATIVES OF ORIENTATION AND COMPOSITIONAL CODING IN V4

Area V4 is the next major stage in the shape-processing hierarchy (Felleman & Van Essen 1991, Gallant et al. 2000, Hansen et al. 2007, Wilkinson et al. 2000). While V1 and V2 are omnibus

areas for all of the visual cortex, supplying information to both dorsal and ventral pathways, V4 is the first area that belongs to the ventral pathway proper. Receptive fields in monkey V4 are larger than in V1 and V2, occupying approximately 1° of visual angle at the fovea and then scaling in diameter roughly 1:1 with distance from the fovea (Gattass et al. 1988).

Within receptive field apertures on this larger scale, the natural world is still mostly smooth, but orientation of 2D boundaries and 3D surface gradients has a strong probability of changing. Because these changes are typically smooth on a local level in the real world, they can be characterized geometrically by curvature, which is the derivative of orientation. V4 neurons combine V1 and V2 inputs to generate information about 2D boundary curvature (Bashivan et al. 2019; Gallant et al. 1993, 1996, 2000; Hegdé & Van Essen 2007; Nandy et al. 2013; Pasupathy & Connor 1999, 2001; Pasupathy et al. 2020; Sharpee et al. 2013; Wilkinson et al. 2000); 3D surface orientation (Hegdé & Van Essen 2005, Hinkle & Connor 2002); 3D surface curvature (Nelissen et al. 2009, Srinath et al. 2021); and possibly higher derivatives, beginning with change in curvature (spirality) (Gallant et al. 1993, 1996).

These transformations compress shape information by summarizing many lower-level signals for oriented contrast. For visible curvature, this summarization is supported by the orientation changing at a predictable rate, which occurs on the level of local shape fragments with high frequency in the natural world. At the extreme of infinite curvature, the change in orientation is a singularity, a 2D or 3D point or a 3D crease. V4 neurons tuned for extremely high or sharp curvature are thus summarizing the conjunction of orientations at the sides of the singularity, with the perceived direction of the point or edge intermediate between the conjoined orientations.

These V4 neurons respond to fragment curvature at specific positions within an infinite variety of larger shapes (Pasupathy & Connor 2002, Srinath et al. 2021). This is the signature characteristic of compositional shape coding, the theory that individual neurons represent the geometry and relative position of shape fragments (parts, geons) (Biederman 1987, Li et al. 2001, Marr & Nishihara 1978, Selfridge 1959). Ensembles of such neurons represent complete shapes as spatial compositions of their constituent geometric fragments. Compositional coding is theoretically appealing because it (*a*) compresses shape information into a handful of geometric fragment signals and (*b*) provides explicit information about the local geometry and overall spatial configuration of those fragments in (*c*) a spatial reference frame centered on and sized to the shape itself, producing stable representation across changes in position and size on the retina.

Compositional codes are also highly productive, in the sense that a small dictionary of parts, e.g., Roman letters, can be combined in different ways to form a vast domain of compositions, e.g., millions of words in Western languages. The domain of imaginable shapes that could be sculpted by physical, biological, or artificial processes is virtually infinite. However, this entire domain, including shapes radically different from any that we have seen before, is immediately perceptible, understandable, and describable if shapes are conceived of by the brain as spatial compositions of familiar geometric fragments.

Figure 1 exemplifies tuning of a single V4 neuron for 2D boundary fragments within larger shape compositions. Across a wide variety of global 2D shapes, this neuron reliably responds to shapes with a sharp convex projection at the top and pointing upward (**Figure 1*a***), corresponding to a tuning peak in curvature and position space (**Figure 1*b***). (Orientation of boundary fragments is a third critical tuning dimension in V4, but in this experiment, it is conflated with angular position.) It might be thought that this position tuning relates to a tiny hot spot in the V4 receptive field, but responsiveness to a particular curvature at a particular object-centered position, e.g., a sharp point at the upper right of an object (**Figure 2**, showing a different neuron from **Figure 1**), is maintained over a spatial range of object positions that makes this

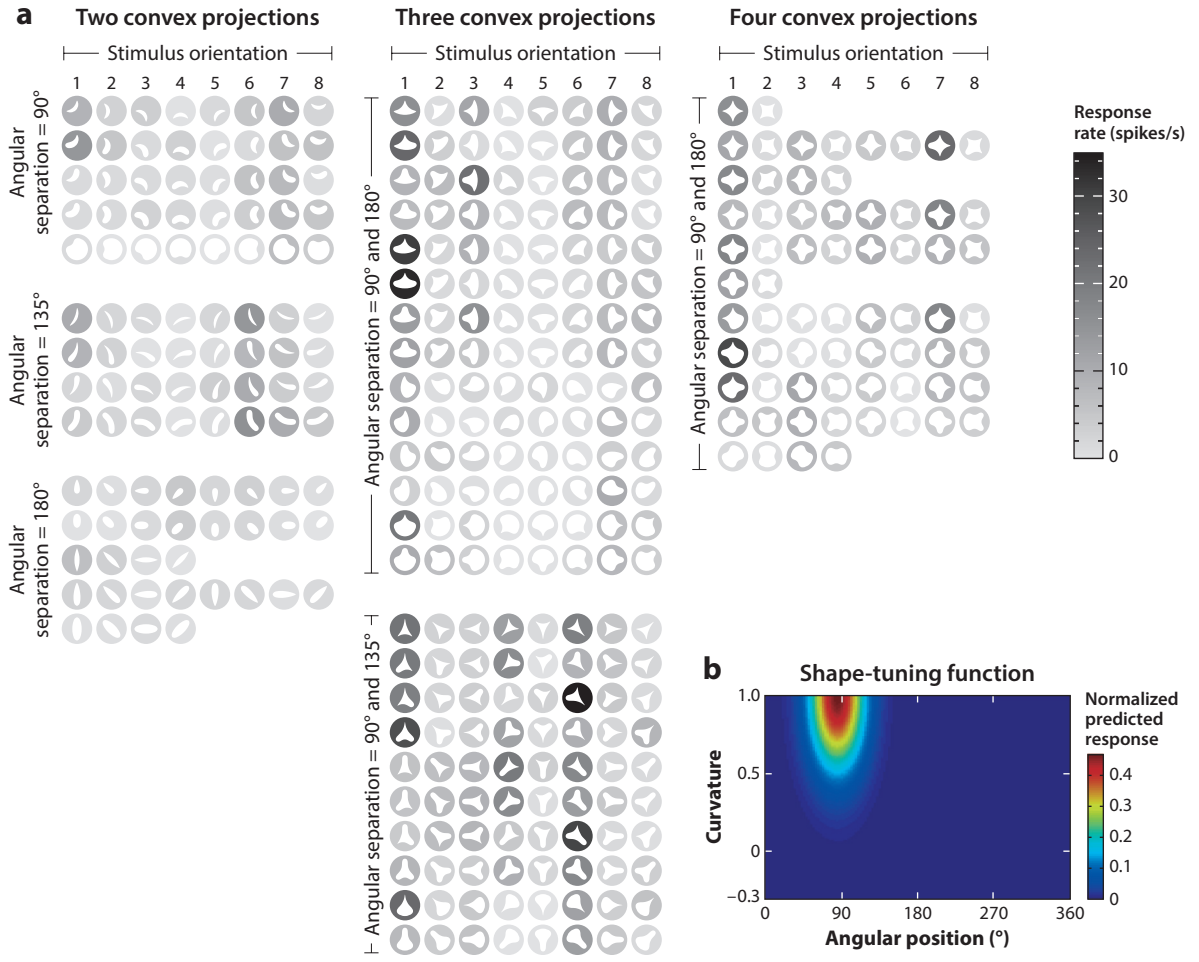


Figure 1

Example V4 neuron shape tuning. (*a*) Responses of an individual V4 neuron are represented by gray levels surrounding each stimulus icon. The scale bar shows that mean response rates ranged from 0 (*light background*) to 34 (*dark background*) spikes/s. The stimulus set comprised most of the geometrically feasible combinations of five standard boundary fragments: sharp convex, medium convex, broad convex, broad concave, and medium concave curves. Each combination was presented at eight orientations (*rows*), or fewer if rotational symmetry made some orientations redundant. The stimuli are arranged into three large blocks (*left*, *middle*, and *right*) according to how many convex projections they contained (two, three, or four). They are also blocked in the vertical direction according to the angular separations between convex projections. The stimuli were presented in red (the optimal color for this cell) at the cell's receptive field center (0.32° left of and 1.32° below fixation). (*b*) A Gaussian shape-tuning function describing the response pattern in panel *a*. The vertical axis represents boundary curvature, and the horizontal axis represents the angular position of boundary fragments with respect to the shape's center of mass. The tuning peak corresponds to sharp convex curvature (1.0) near the top of the shape (84.6°). Figure adapted with permission from Pasupathy & Connor (2002).

explanation untenable (**Figure 2*a***). Tuning for shapes with this point at the upper right (**Figure 2*a***, left), versus similar shapes without this point (**Figure 2*a***, right), is maintained across the V4 receptive field. At every position where this neuron is responsive to the point at the upper right (**Figure 2*a***, left), it is nonresponsive to the same shape without the point (**Figure 2*a***, right). The neuron is sensitive to retinal stimulus position only according to the typical 2D Gaussian profile of V4 receptive field sensitivity. (In these miniature stimulus position icons, the size of the

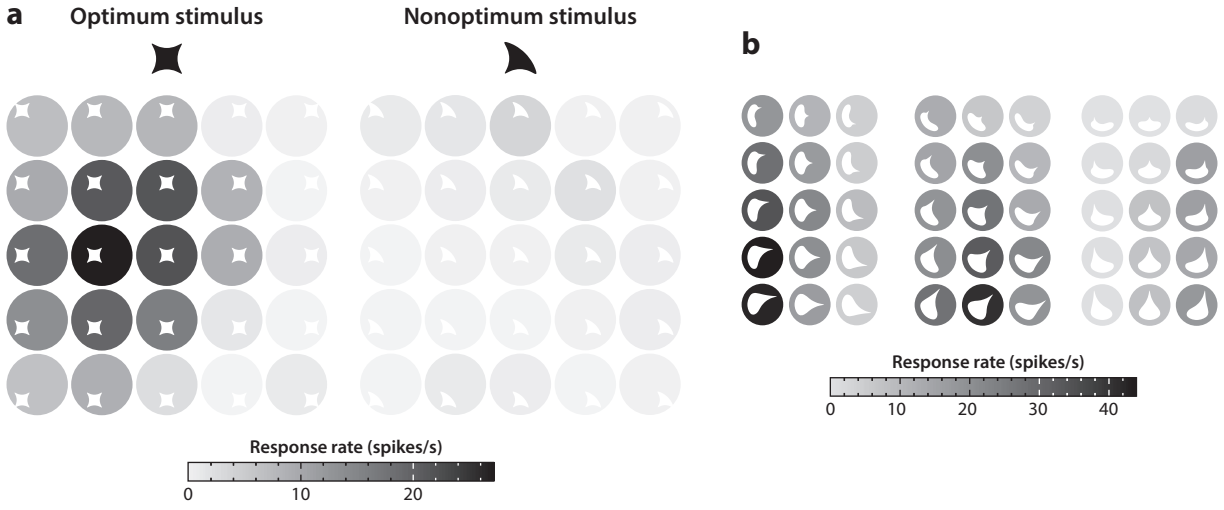


Figure 2

V4 neural tuning for spatial position. (*a*) Consistency across absolute spatial position. Surrounding gray levels denote average responses to an optimum (*left*) and nonoptimum (*right*) stimulus, both shown in black at the top, presented at 25 positions on a 5 × 5 grid centered on the receptive field center. In this plot, the background circles are larger than the estimated receptive field. Shape tuning for the optimum shape, which contains a sharp convexity pointing to the upper right, versus the nonoptimum shape remains consistent out to the limits of the receptive field. (*b*) Acute sensitivity to object-relative spatial position. Stimuli are on the same order of size as in panel *a* but are shown with larger icons to clarify small shape differences. The orientation and relative position of the sharply pointed convex projection that drove this neuron's responses were systematically varied, as shown by the stimulus icons. Response rates are indicated by the surrounding gray levels. The neuron was acutely tuned for convexity positioned near the upper right of the object, as well as the orientation of that convexity (compare responses to three blocks with orientation of the sharp point at 0° on the left, 45° in the center, and 90° on the right). Figure adapted with permission from Pasupathy & Connor (2001).

background circles relative to the stimulus icons is larger than the estimated receptive field extent.) In contrast, this neuron exhibits sharp tuning for object-centered spatial position on the fine scale of the stimulus itself (**Figure 2*b***). **Figure 2*b*** also illustrates tuning for orientation of the sharp point (compare the three blocks).

At the population level, an entire shape (**Figure 3*a***), from the stimulus set in **Figure 1**, evokes peaks in geometry and position space corresponding to its constituent fragments (**Figure 3*b***), and the shape can be reconstructed from that pattern (**Figure 3*c***). A comparable population plot for V1 would have all of the response energy concentrated near 0 curvature. V4 neurons summarize many V1 orientation signals with tuning for curvature, orientation, and object-relative position. This transformation compresses the V1 orientation-based representation approximately eightfold, partly by emphasizing sharp curvature signals (Carlson et al. 2011), which carry more shape information (Attneave 1954). It provides the first explicit signals for shape parts that are large enough to be perceptible and nameable (as in the sharp spikes, concave indentations, etc. in **Figure 3**), and it confers stability on the scale of V4 receptive fields (**Figure 2**). This stability across a larger area of space does not represent a loss of spatial information. Instead, it is a transformation of retinotopic spatial information into relative spatial information. Relative spatial information is carried by signals for geometric derivatives (curvature and higher derivatives like spirality) (Connor & Knierim 2017; Gallant et al. 1996, 2000; Wilkinson et al. 2000). These derivatives summarize spatial patterns (curves and angles). Relative spatial information is also carried by neural tuning for the object-relative position of geometric fragments (like curves and angles). Because of this transformation to relative spatial coordinates, we have only coarse cognitive access to retinotopic

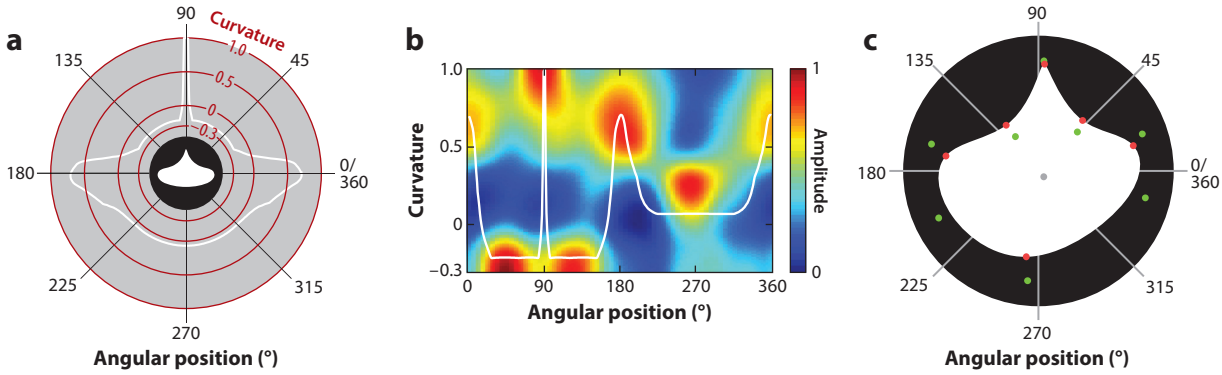


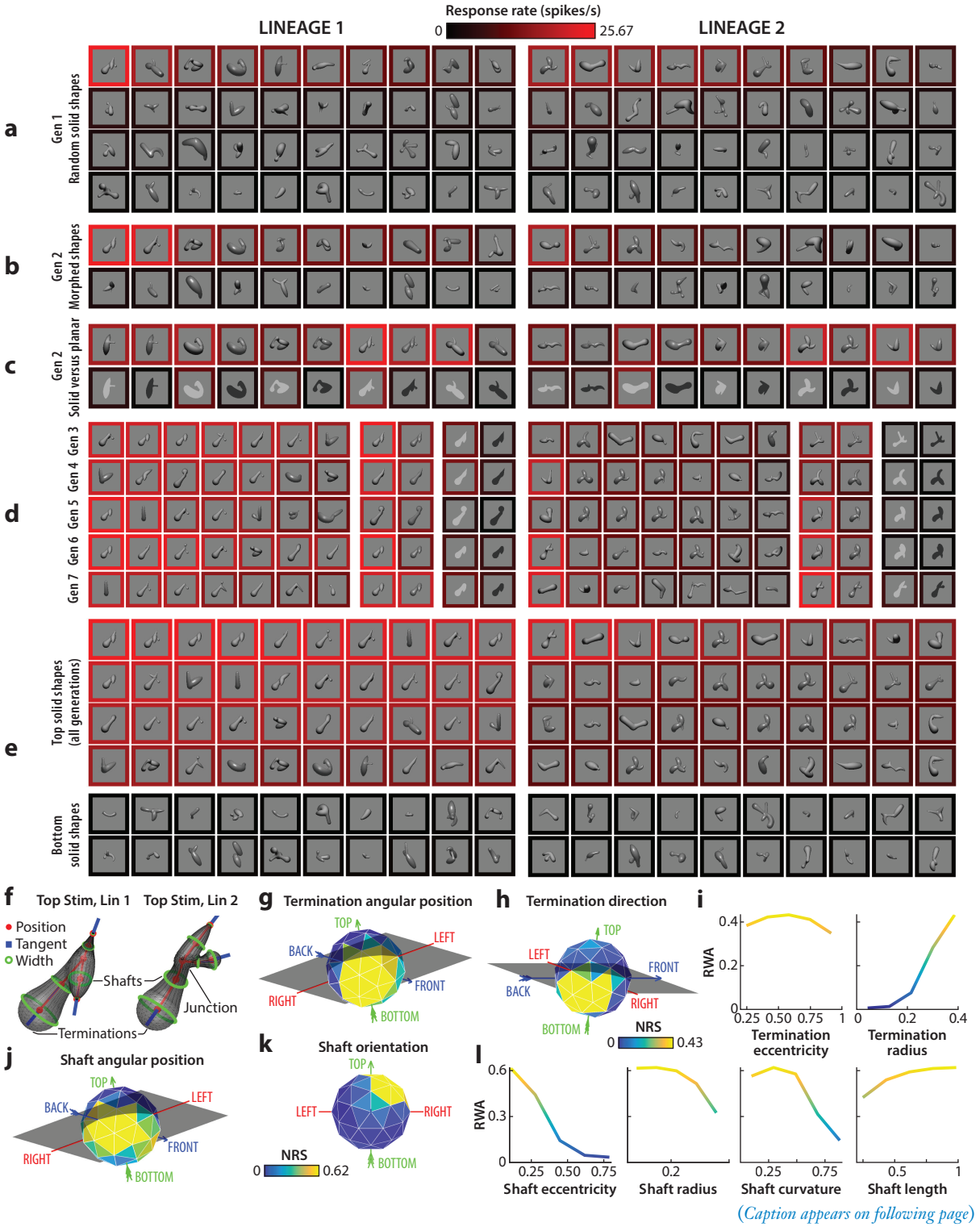
Figure 3

A population response to an example shape in V4. (*a*) The white line represents this shape's boundary curvature as a function of angular position, shown in a polar format centered around the shape's center of mass to highlight the correspondence with boundary features. (*b*) The estimated population response across the curvature × position domain (*colored surface*) with the veridical curvature function superimposed (*white line*). A Cartesian plot is used because a polar plot would distort peak width in the population response. The surface was thresholded by subtracting the minimum value from all points, so amplitude varies from 0 (minimum) to 1 (maximum). (*c*) A reconstruction of the geometric shape from the population surface in panel *b*. Figure adapted with permission from Pasupathy & Connor (2002).

spatial information but an acute sense of local spatial relationships, in the sense of curvature perception (Koenderink 1993, Koenderink et al. 1997, Watanabe et al. 1999) and understanding the spatial compositions of complete shapes.

While some V4 neurons encode 2D shape fragments in this way, others encode 3D shape fragments. Both domains are critical kinds of shape perception, 2D for symbols and patterns on surfaces and 3D for volumetric object structure. Coding for 3D shape fragments is exemplified in **Figure 4**, where separate lineages of evolving shapes, guided by the neuron's response levels (**Figure 4*a–e***), lead to shapes characterized by a long projection pointing toward the lower left and ending in a bulbous termination (**Figure 4*f***). This selectivity can again be quantified in terms of local geometry (medial axis shape of the projecting limb and its termination, 3D surface orientations and surface curvatures) and 3D object-relative position (**Figure 4*g–l***). Thus, V4 generates a compressed, stable representation of perceptible, nameable 2D and 3D shape fragments. The two domains appear to be segregated across the surface of V4 (**Figure 5**), suggesting their independent importance and the need for separate addressability by input and output connections.

Artificial neural networks have become increasingly important for modeling and interpreting neural tuning in the ventral pathway (e.g., Cadena et al. 2024, Pospisil et al. 2018, Yamins et al. 2014). By multiple measures, V4 seems most homologous to layer 3 in artificial networks with depths (numbers of layers) comparable to the ventral pathway (five or six convolutional layers that tile visual space with repeated copies of the same filters). In particular, models based on linear combinations of layer 3 response patterns do the best job of predicting response patterns of V4 neurons (Yamins et al. 2014) relative to other layers. Remarkably, individual layer 3 neurons exhibit V4-like tuning for 2D or 3D shape fragments in an object-centered reference frame (Pospisil et al. 2018, Srinath et al. 2021). As in V4, layer 3 shape coding is biased toward 3D volumetric shape (Srinath et al. 2021). Visual images designed to evoke maximum responses from artificial network models of individual V4 neurons, sometimes called superstimuli, also show a remarkable tendency toward 3D shape-from-shading patterns, combined with texture-like repetition of these motifs (Bashivan et al. 2019).



(*Caption appears on following page*)

Figure 4 (*Figure appears on preceding page*)

Example V4 neuron encoding solid 3D shape information. (*a*) The first generation (Gen 1) of 40 random stimuli per lineage. Each stimulus was rendered with a lighting model based on either a matte or polished surface. Stimuli were centered on and sized to fit within the previously mapped receptive field of an individual V4 neuron while the monkey performed a fixation task. The neuron's average response to each stimulus is represented by the color of the surrounding border, referenced to the scale at the top, with bright red corresponding to 25.67 spikes/s. Stimuli in each block are ordered by descending response strength from the upper left to the lower right. (*b*) Half of Gen 2 comprised partially morphed descendants of ancestor stimuli from Gen 1 plus additional random stimuli. (*c*) The other half of Gen 2 comprised tests of high-response Gen 1 stimuli rendered as solid versus flat shapes. (*d*) The highest response stimuli and example solid–flat comparisons in Gen 3–7. (*e*) The highest and lowest response stimuli across all generations. (*f*) Parameterization of shaft, junction, and termination shape, illustrated for the top stimuli (Stim) for Lineages (Lin) 1 and 2. (*g–l*) An RWA analysis of response strength. Each panel shows the average normalized response strength as a function of geometric dimensions used to describe shaft or termination shape. Each plot represents a slice through the RWA at the location of the overall RWA peak across all dimensions (rather than a collapsed average across the other dimensions). Spherical (object-centered position and termination direction) and hemispherical (shaft orientation) dimensions are shown as spherical polygons, in some cases tilted and rotated to reveal the tuning peak. The arrows and labels (*LEFT*, *RIGHT*, *TOP*, *BOTTOM*, *BACK*, and *FRONT*) indicate the original directions in the stimulus from the monkey's point of view. Normalized response strength is indexed to a color scale for shafts (below panel *k*) and a color scale for terminations (below panel *h*). Color is a redundant cue for response strength in the Cartesian plots. Abbreviations: NRS, normalized response strength; RWA, response weighted average. Figure adapted with permission from Srinath et al. (2021).

FRAGMENT COMBINATIONS IN THE INFEROTEMPORAL CORTEX

Signals from V4 feed into the inferotemporal cortex (IT), which comprises approximately four subsequent processing stages along its posterior to anterior extent (Bao et al. 2020). It has long been known that IT neural responses to complex shapes (Gross et al. 1972) are actually driven by parts of those shapes (Tanaka 1996, Tanaka et al. 1991), consistent with compositional coding, as described above for V4. In IT, however, these shape fragments are more complex (Kobatake & Tanaka 1994, Tanaka et al. 1991), combining multiple simpler fragments of the type encoded in V4, both 2D (Brincat & Connor 2004) and 3D (Hung et al. 2012, Yamane et al. 2008; see also Janssen et al. 1999, 2000a,b). These multiple fragments can again be described in terms of their geometry and object-relative positions (**Figure 6**). Tuning for fragment combinations is remarkably stable across changes in depth cues (**Figure 6*c***), lighting direction (**Figure 6*d***), stereoscopic depth (**Figure 6*e***), spatial position (**Figure 6*f***), and stimulus size (**Figure 6*b***).

This combined sensitivity to multiple geometric fragments varies across neurons from linear summation (response to shapes with fragments A and B = response to shapes with A only + response to shapes with B only) to nonlinear selectivity (strong responses only to shapes with both fragments A and B) (Brincat & Connor 2006), exemplified in **Figure 6*i*** (the equation shows that parts A and B contribute less than 1 spike/s to the response, while the combination A + B, represented by the product term, has a weight of 49 spikes/s). Nonlinear responses evolve across time, with linear summation appearing early, many neurons transitioning from linear to nonlinear coding, and purely nonlinear neurons responding later. As a result, nonlinear signals peak approximately 60 ms later than linear signals, suggesting that they depend on recurrent network processing (Brincat & Connor 2006).

This expenditure of neural processing time has an important purpose, because only nonlinear processing produces explicit information about fragment combinations. Linear summation generates confusing signals at the level of individual neurons. Imagine a neuron that sums information about two fragments, A and B. High-contrast shapes with fragment A alone could evoke responses comparable to low-contrast shapes with both A and B. Definitive information about fragment combinations remains implicit, available only at the population level, and requiring further decoding. Now imagine a neuron with nonlinear selectivity, with no response to fragment A or B alone but a strong response to shapes containing both fragments A and B. Neurons like this

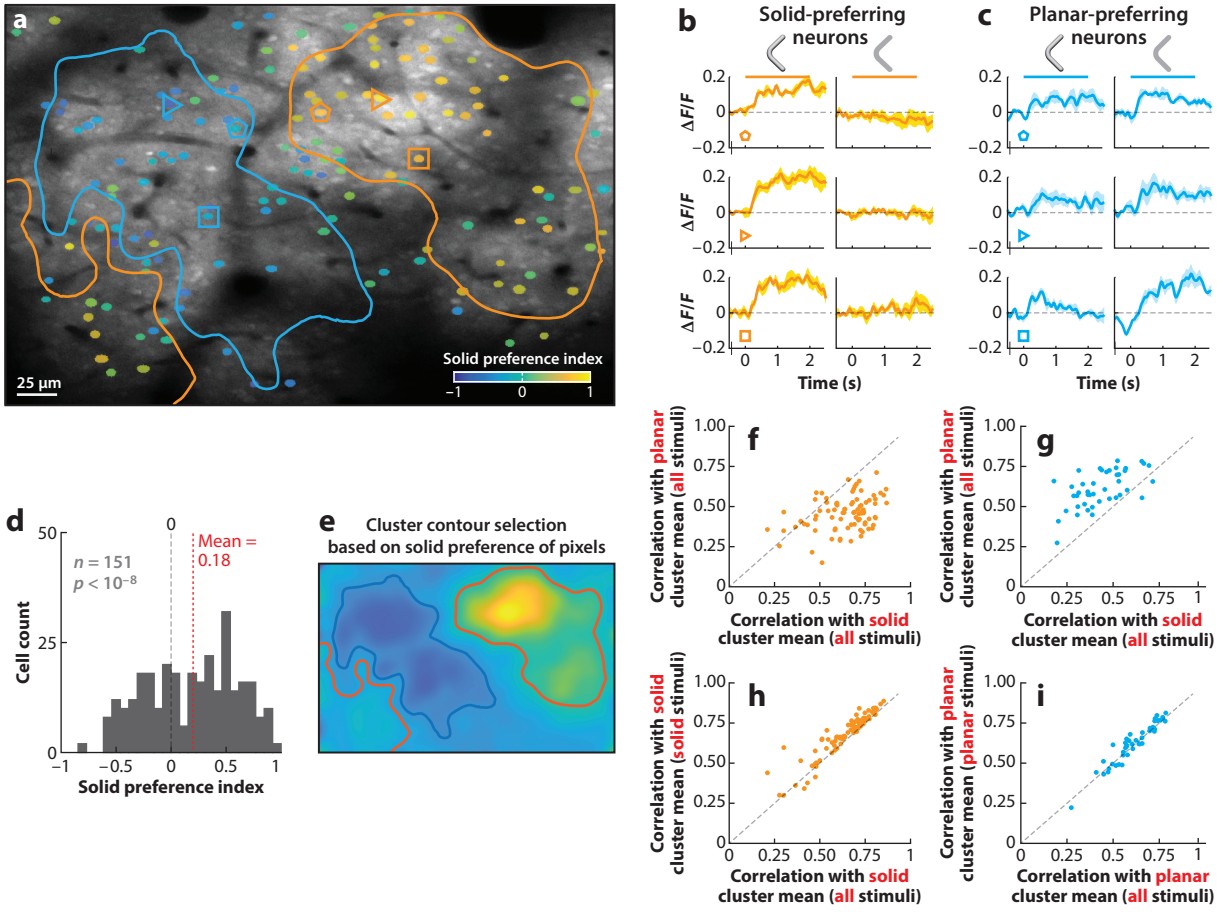


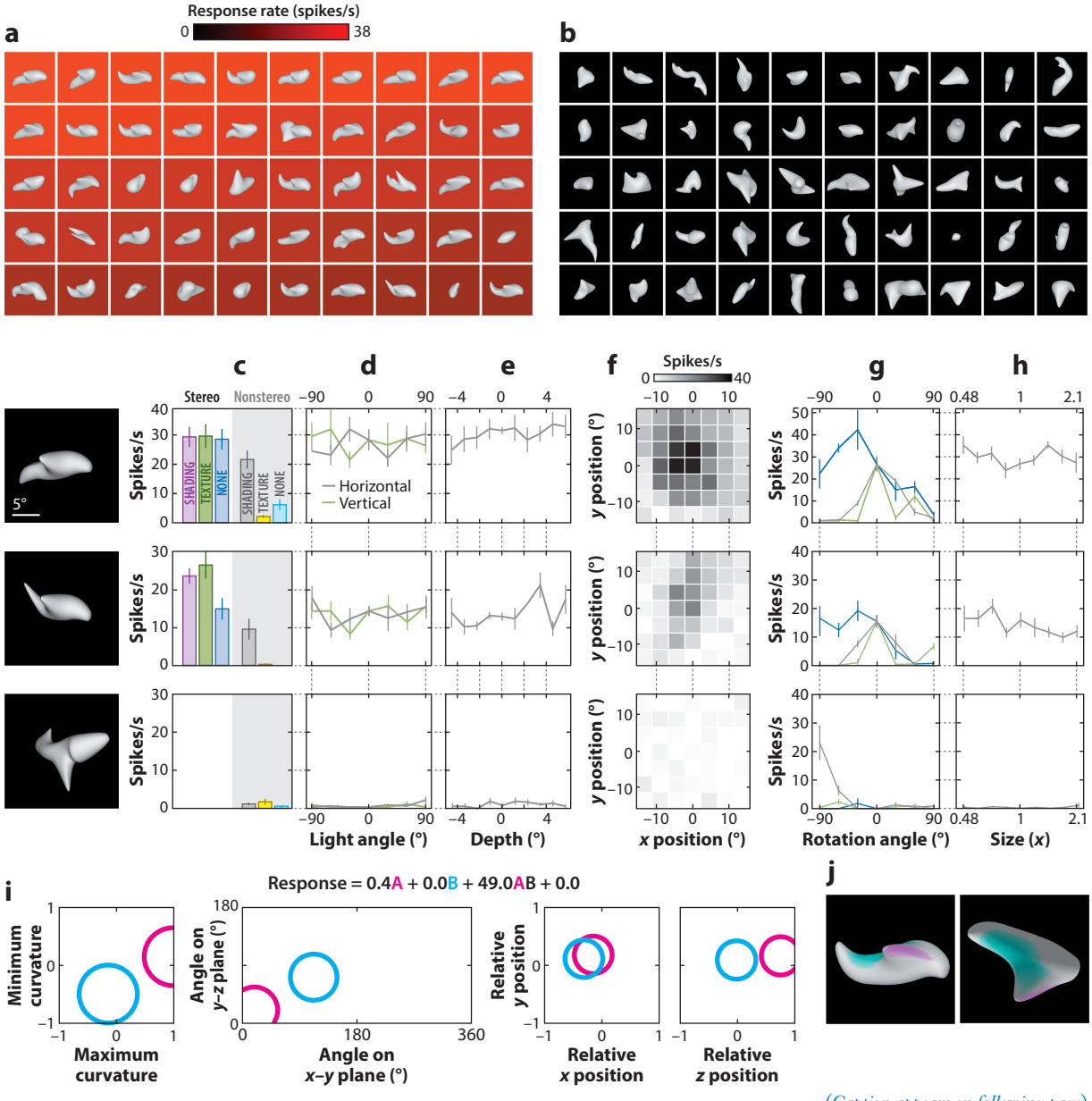
Figure 5

Micro-organization of 2D flat and 3D solid shape coding. (*a*) An anatomical average image of a section of the V4 cortical surface (anatomical scale bar at *lower left*). Neurons with significant responses to stimuli, as measured by changes in fluorescence of the calcium indicator Oregon Green BAPTA-1AM, are overlaid with a color indexing their preference for 2D flat (*blue*) or 3D solid (*yellow*) stimuli (see the solid preference scale bar at *lower right*). (*b*) Peristimulus time fluorescence plots for the example 3D solid-preferring neurons indicated by orange polygons in panel *a*. The horizontal bars span the 2 s stimulus presentation period for solid (*left*) and flat (*right*) stimuli. (*c*) Example 2D flat-preferring neurons; details are as in panel *b*. (*d*) Distribution of solid preference values for neurons in this region. The mean solid preference value of 0.18 is significantly greater than 0 (*t*-test, two-tailed, $p < 10^{-6}$). (*e*) A smoothed map of solid preference values in this imaging region, used as the basis for drawing the cluster boundaries shown in panel *a*. (*f*) Correlations of stimulus response patterns for neurons in the upper right solid cluster (*orange contour* in panel *a*) with the average response pattern in that solid cluster (*horizontal axis*) versus the average response pattern in the planar cluster (*blue contour* in panel *a*). (*g*) Correlations of stimulus response patterns for neurons in the planar cluster (*blue contour* in panel *a*) with solid and planar cluster averages. (*h*) Correlations of solid cluster neurons with the solid cluster average across all stimuli (*horizontal axis*) versus solid stimuli only (*vertical axis*). (*i*) Correlations of planar cluster neurons with the planar cluster average across all stimuli (*horizontal axis*) versus planar stimuli only (*vertical axis*). Figure adapted with permission from Srinath et al. (2021).

provide explicit signals (not requiring further decoding) for complex shape configurations, like the dorsal torso surface + horizontal fin in **Figure 6**.

Compositional coding in IT, in terms of more complex parts comprising multiple V4-like geometric fragments, represents another step in the compression of shape information. Signals for more complex parts convey more geometric information, summarizing across higher fractions

of object boundaries or surfaces, and thus, representation of entire shapes requires fewer signals. At the same time, the dimensionality of the parts domain increases, presumably requiring more IT neurons (a limited resource) to represent the entirety of that domain in detail. The solution to this dilemma may be that IT neurons selectively develop tuning for fragment combinations based on experience in the natural world. It has been shown that extensive training of monkeys on tasks depending on geometric combinations produces increased tuning for those combinations (Baker et al. 2002, Cheng et al. 2024). This might represent a more general process that occurs



(*Caption appears on following page*)

Figure 6 (*Figure appears on preceding page*)

Example inferotemporal cortex (IT) neuron tuned for a configuration of surface shape fragments. Error bars indicate standard error of the mean (SEM). (*a*) The top 50 stimuli across 8 generations (400 stimuli) for a single IT neuron recorded from the ventral bank of the superior temporal sulcus. (*b*) The bottom 50 stimuli for the same cell. (*c*) Responses to highly effective (*top*), moderately effective (*middle*), and ineffective (*bottom*) example stimuli as a function of depth cues (shading, disparity, and texture gradients). Responses remained strong as long as disparity (*purple*, *green*, and *dark blue*) or shading (*gray*) cues were present. The cell did not respond to stimuli with only texture cues (*yellow*) or silhouettes with no depth cues (*light blue*). (*d*) The response consistency across lighting direction. The implicit direction of a point source at infinity was varied across 180° in the horizontal (left to right; *grey curve*) and vertical (below to above; *green curve*) directions, creating very different 2D shading patterns. (*e*) The response consistency across stereoscopic depth, varying the disparity of the surface point at fixation from −4.5° (near) to 5.6° (far). (*f*) The response consistency across *x* and *y* positions. (*g*) The sensitivity to stimulus orientation. Like all neurons in our sample, this cell was highly sensitive to stimulus orientation, although it showed broad tolerance (approximately 90°) to rotation about the *z* axis (rotation in the image plane; *blue curve*); this rotation tolerance is also apparent among the top 50 stimuli in panel *a*. Rotation out of the image plane, about the *x* axis (*grey*) or *y* axis (*green*), strongly suppressed responses. (*h*) The response consistency across object size over a range from half to twice the original stimulus. (*i*) A linear–nonlinear response model based on two Gaussian tuning functions. The Gaussian functions describe tuning for surface fragment geometry, defined in terms of curvature (principal, i.e., maximum and minimum, cross-sectional curvatures), orientation (of a surface normal vector, projected onto the *x*–*y* and *y*–*z* planes), and position (relative to the object's center of mass in *x*–*y*–*z* coordinates). The curvature scale is squashed to a range between –1 and 1. The 1.0 standard deviation boundaries of the two Gaussians (*magenta* and *cyan*) are shown projected onto different combinations of these dimensions. The equation shows the overall response model, with fitted weights for the two Gaussians, the product or interaction term, and the baseline response. (*j*) The tuning functions are projected onto the surface of a high response stimulus, seen from the observer's viewpoint (*left*) and from above (*right*). Figure adapted from Yamane et al. (2008).

throughout life, producing stronger representation of naturally common geometric combinations, like the fin–torso combination in **Figure 6**.

This process in IT of combining simpler parts into more complex configurations could theoretically continue, even to the point of combinations equivalent to entire shapes. However, based on currently available data, the process does not continue past the point of combining two to three geometric fragments of the type shown in **Figure 6** (Yamane et al. 2008), as well as fragments best described in terms of medial axis shape (Hung et al. 2012). This is true even after long-term training in match-to-sample tasks with relatively simple, 2D letter-like stimuli. This training produces increased selectivity for combinations of letter parts, which presumably enhances the perception of learned letters, and even transfers to the same part combinations in unfamiliar letters. However, it does not produce exclusive selectivity for single letters (Cheng et al. 2024). Categorical identity information remains implicit within, although linearly decodable from, the monkey visual cortex (Hung et al. 2005, Kriegeskorte et al. 2008, Sigala & Logothetis 2002, Vogels 1999).

The original description of neurons selective for single persons, i.e., grandmother (Gross 2002) or Jennifer Anniston cells, in the human medial temporal lobe (Quian Quiroga et al. 2005) was revised to reflect that these neurons, found in memory-related structures like the hippocampus and amygdala, respond to multiple persons and places associated in personal memory (Quian Quiroga et al. 2008). Signals for individual category identities appear in the prefrontal cortex when they must be preserved in the short-term memory (Freedman et al. 2001) but not when categorical identity is only an implicit, transitional task variable (Sasikumar et al. 2018). Categorical identity appears to be implicit until needed.

The obvious reason for limits on part-coding complexity is the combinatorial explosion of neurons that would be required for 1:1 representation (the asymptote of compression) of all possible shapes in an unpredictable world of very complex shapes. Even at the complexity level described above, the number of distinguishable fragment combinations that must be encoded is at least on the order of 10^3 (Tanaka 1993). However, in addition to this mathematical constraint on 1:1 coding, the flexibility of parts-level compositional coding may add another kind of stability, an inherent robustness for recognition of familiar shapes under partial occlusion. This is a common

viewing condition in crowded natural scenes and a particular vulnerability for artificial vision networks, which report object category probabilities with 1:1 coding in a final output layer (Fawzi & Frossard 2016, Kortylewski et al. 2021, Zhu et al. 2019).

Stability under partial occlusion may be a consequence of how category probabilities change as a function of the number of visible parts. Given the highly specific information about part geometry and spatial composition encoded by IT neurons, and under reasonable assumptions about numbers of parts in a shape and numbers of parts in the neural dictionary, recognition requires only two to three parts for near-certain identification (Cheng et al. 2024). This provides a parsimonious explanation for the extraordinary stability of familiar shape recognition under frequently crowded viewing conditions in which only a fraction of a given shape is visible. It does not matter which fraction is visible for a compositional coding system that can identify a familiar shape based on any subset of shape parts.

ORIENTATION STABILITY IN THE INFEROTEMPORAL CORTEX

Compositional shape representation in an object-relative reference frame confers stability for size and position (see above). This stability is extended across the larger receptive fields in IT (Brincat & Connor 2004, Ito et al. 1995, Yamane et al. 2008). However, compositional coding does not confer stability across changes in object orientation. This is a particularly limiting problem for 3D objects, which can rotate relative to the frontoparallel plane of retinotopic vision, potentially changing the entire set of object parts that are visible versus hidden by self-occlusion.

For both familiar and unfamiliar objects, there is some stability of shape perception across views, on the order of 30–60° for rotations out of the frontoparallel plane (Hung et al. 2012, Tanaka et al. 1991) and up to near 90° for rotations in the frontoparallel plane (Hung et al. 2012). Recognition of familiar objects across wider orientation ranges depends on learning associations between views (Logothetis & Pauls 1995, Wang et al. 2005, Yamaguchi et al. 2016, Zhao et al. 2018), in some cases based on spatiotemporal continuity across different views during natural visual experience (Isik et al. 2012, Li & DiCarlo 2008, Vetter et al. 1995).

It makes sense for neurons to have limited tolerance for orientation changes of objects in the real world, given the many purposes of shape perception beyond recognition. Absolute object orientation, in the gravitational reference frame of the world, affects our physical understanding of and interaction with even familiar objects. An upside-down coffee cup presents very different physical affordances from an upright cup. However, orientation changes of the eyes themselves are problematic even in the gravitational reference frame. Vision is physically based on a platform, the head, that constantly rotates with respect to gravity during normal physical activity. Changes in elevation (pitch) and azimuth (yaw) of the head are tightly compensated for by counter-rotations of the eyes based on vestibulo-ocular reflexes. However, rotations around the line of sight (roll or lateral tilt) are only weakly compensated for by countertorsional eye movements, on the order of approximately 20% (Miller 1962, Schworm et al. 2002). This makes retinotopic images problematically unstable with respect to the gravitational world. The surface orientation of the ground plane on which objects rest, the orientation of structures orthogonal to this plane (like upright plants and animals), the balance of structures sitting on this plane, and the movement potentials of objects in the gravitational field are all critical to physical understanding and behavioral planning. Thus, the continual dissociation of retinal image orientation from the gravitational reference frame, by 90° or more, requires some compensatory mechanism in the visual cortex.

This compensatory mechanism, in central and anterior IT, is transformation of retinal image information into a gravitational reference frame that remains stable across head rotations. A majority of neurons in monkey IT respond to 3D shape stimuli based on their orientation

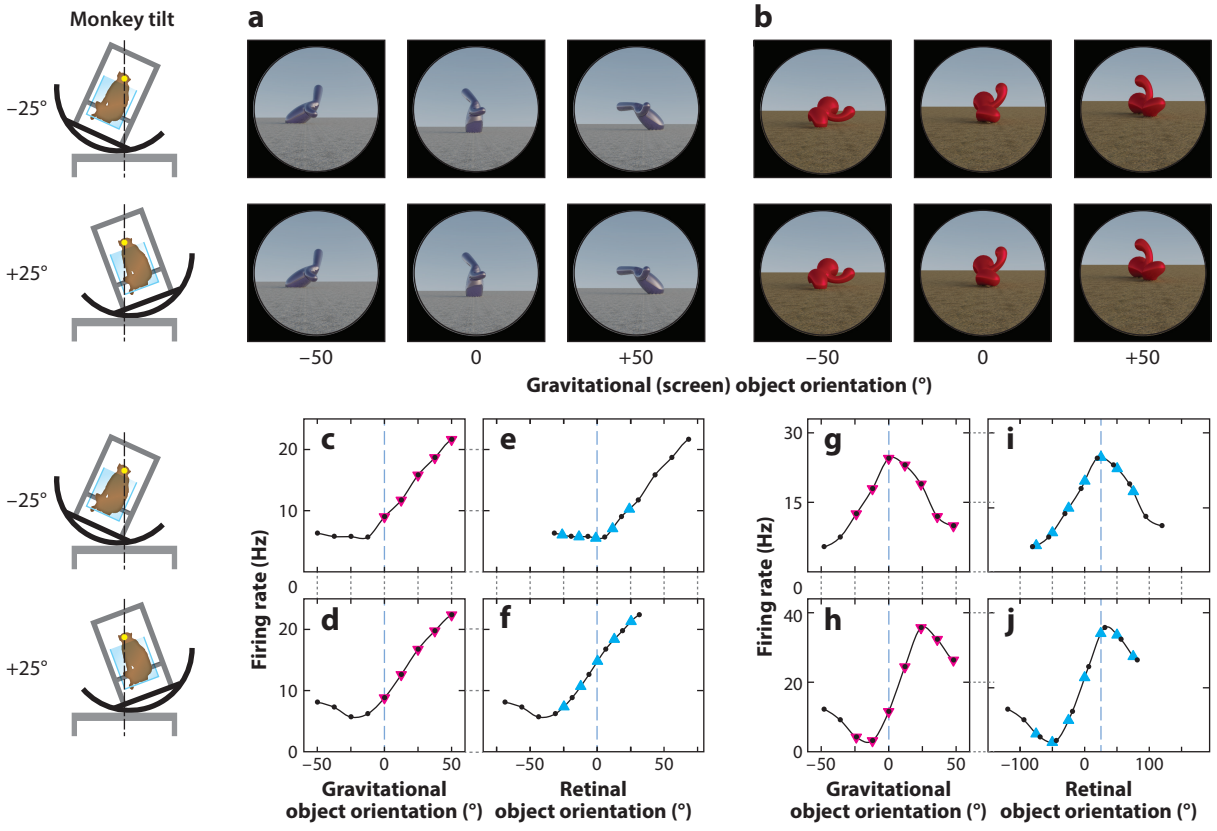


Figure 7

Example neurons tuned in gravitational space and retinal space. (*a*, *b*) Stimuli demonstrating example object orientations in the full scene condition. The orientation discovered in the genetic algorithm experiments is arbitrarily labeled 0°. The two monkey tilt conditions are diagrammed at the left. The small yellow dots at the center of the head (connected by *vertical dashed lines*) represent the virtual axis of rotation produced by a circular sled supporting the chair. Stimuli were presented on a 100°-wide display screen for 750 ms (separated by 250 ms blank screen intervals) while the monkey fixated a central dot. Stimuli were presented in random order for a total of 5 repetitions each. (*c*, *d*) Responses of a gravitationally tuned IT neuron studied with the stimuli shown in panel *a*, as a function of object orientation on the screen and thus with respect to gravity, across a 100° orientation range, while the monkey was tilted (*c*) –25° and (*d*) 25°. Response values are averaged across the 750-ms presentation time and across 5 repetitions and smoothed with a boxcar kernel of width 50° (3 orientation values). For this neuron, object orientation tuning remained consistent in screen/gravity space across the two tilt conditions. The pink triangles indicate object orientations comparable across tilts in the gravitational reference frame. (*e*, *f*) The same data plotted against orientation on the retina, corrected for 6° counter-rolling of the eyes in each tilt condition. The cyan triangles indicate response values comparable across tilts in the retinal analysis. Due to the shift produced by ocular counter-rolling, these comparable values were interpolated between tested screen orientations using a Catmull-Rom spline. Since orientation tuning was consistent in gravitational space, the tuning functions in retinal space are shifted right or left by about 20° each. (*g*, *h*) Responses of a retinally tuned IT neuron studied with the stimuli shown in panel *b*, as a function of object orientation on the screen and thus with respect to gravity, across a 100° orientation range, while the monkey was tilted (*g*) –25° and (*h*) 25°. In this case, the tuning peak was shifted about 40° in the direction expected for orientation tuning in retinal space. (*i*, *j*) The same data plotted against orientation on the retina, corrected for 6° counter-rolling of the eyes in each tilt condition. The correspondence between curves in panels *i* and *j*, with peaks at near 0°, is consistent with orientation tuning in retinal space. Figure adapted with permission from Emonds et al. (2023) (CC BY 4.0).

in the gravitational reference frame, across at least a 50° range of lateral head and body tilts (Emonds et al. 2023; see also Rosenberg et al. 2013) (**Figure 7**). This appears to be partly based on vestibular or somatosensory cues for head or body orientation but also appears to be based on visual cues for self-orientation with respect to the gravitational environment, which are

extensively processed in the middle channel of anterior IT (Vaziri & Connor 2016, Vaziri et al. 2014). Other IT neurons respond to 3D shape stimuli based on their tilted representation on the retina. This parallel representation of retinal and gravitational reference frames is consistent with our ability to access either reference frame voluntarily (Attneave & Reid 1968). Both are important under the many conditions in which we are not consistently upright (resting, reaching, participating in sports, working in cramped environments) (Halberstadt & Saitta 1987, Krumhuber et al. 2007, Mara & Appel 2015, Mignault & Chaudhuri 2003, Zikovitz & Harris 1999). The retinal–head–body reference frame is more relevant to how we grasp an object in these conditions, while the gravitational reference frame determines the direction in which we lift or throw it.

This result addresses a longstanding question about the reference frame for shape perception. Evidence cited above has shown that the position and size of this reference frame is determined by the object itself, a transformation from the original coordinate system of the retina. What is the orientation of this object-based reference frame? It has been proposed that even the orientation of shape processing could be determined by the object itself, at least for familiar objects that have a canonical orientation (Farah & Hammond 1988, Farah et al. 1994). In contrast, based on results discussed above, when eye, head, body, and gravitational orientation are aligned, ventral pathway responses to unfamiliar objects are predictable in a spatial reference frame aligned with these egocentric and allocentric axes (see above). However, when egocentric and allocentric axes diverge, IT neurons maintain compositional coding in both egocentric (retinal) and allocentric (gravitational) reference frames, in separate populations, optimizing information necessary for divergent purposes.

CONCLUSION

Shape processing consumes a substantial fraction of the brain's computational resources. It is implemented across many stages, beginning in the retina and extending to the high-level visual cortex in the ventral pathway (as reviewed in this article), as well as the dorsal pathway (Rosenberg et al. 2023). It is an automatic, ongoing component of conscious experience and understanding of the world. It makes the spatial, physical, mechanical, and biological structure of the world and the things within it intelligible to us.

In this review, we suggest an explanatory framework for understanding how the many successive stages of shape processing contribute to this intelligibility. Vision begins with the massive, confusing, constantly variable information stream from photoreceptors in the retina. In a stepwise fashion, this information is compressed from megapixels to what could be thought of as a handful of geometric fragments, mentally arranged in a local reference frame and explicitly representing the kind of shape information that we perceive and use in a fashion that remains stable across viewing conditions. Somehow, at some point in this progression toward increasingly intelligible neural information, awareness and understanding of shape emerge.

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.

LITERATURE CITED

Adams DL, Horton JC. 2003. A precise retinotopic map of primate striate cortex generated from the representation of angioscotomas. *J. Neurosci.* 23:3771–89

Afraz S-R, Kiani R, Esteky H. 2006. Microstimulation of inferotemporal cortex influences face categorization. *Nature* 442:692–95

Attneave F. 1954. Some informational aspects of visual perception. *Psychol. Rev.* 61:183–93
Attneave F, Reid KW. 1968. Voluntary control of frame of reference and slope equivalence under head rotation. *J. Exp. Psychol.* 78:153–59
Baker C, Behrmann M, Olson C. 2002. Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nat. Neurosci.* 5:1210–16
Bao P, She L, McGill M, Tsao DY. 2020. A map of object space in primate inferotemporal cortex. *Nature* 583:103–8
Barlow HB. 1961. Possible principles underlying the transformation of sensory messages. In *Sensory Communication*, ed. WA Rosenblith, pp. 217–33. Cambridge, MA: MIT Press
Bashivan P, Kar K, DiCarlo JJ. 2019. Neural population control via deep image synthesis. *Science* 364:eaav9436
Biederman I. 1987. Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94:115–47
Brincat SL, Connor CE. 2004. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nat. Neurosci.* 7:880–86
Brincat SL, Connor CE. 2006. Dynamic shape synthesis in posterior inferotemporal cortex. *Neuron* 49:17–24
Cadena SA, Willeke KF, Restivo K, Denfield G, Sinz FH, et al. 2024. Diverse task-driven modeling of macaque V4 reveals functional specialization towards semantic tasks. *PLOS Comput. Biol.* 20(5):e1012056
Carlson ET, Rasquinha RJ, Zhang K, Connor CE. 2011. A sparse object coding scheme in area V4. *Curr. Biol.* 21:288–93
Cheng A, Sokol S, Connor CE. 2024. An inferotemporal coding strategy robust to partial object occlusion. bioRxiv 2024.04.09.588746. **https://doi.org/10.1101/2024.04.09.588746**
Connor CE, Knierim JJ. 2017. Integration of objects and space in perception and memory. *Nat. Neurosci.* 20:1493–503
Cumming BG, Parker AJ. 1999. Binocular neurons in V1 of awake monkeys are selective for absolute, not relative, disparity. *J. Neurosci.* 19:5602–18
Curcio CA, Allen KA. 1990. Topography of ganglion cells in human retina. *J. Comp. Neurol.* 300:5–25
Curcio CA, Sloan KR, Packer O, Hendrickson AE, Kalina RE. 1987. Distribution of cones in human and monkey retina: individual variability and radial asymmetry. *Science* 236:579–82
De Valois RL, Albrecht DG, Thorell LG. 1982. Spatial frequency selectivity of cells in macaque visual cortex. *Vis. Res.* 22:545–59
Emonds AMX, Srinath R, Nielsen KJ, Connor CE. 2023. Object representation in a gravitational reference frame. *eLife* 12:e81701
Farah MJ, Hammond KM. 1988. Mental rotation and orientation-invariant object recognition: dissociable processes. *Cognition* 29:29–46
Farah MJ, Rochlin R, Klein KL. 1994. Orientation invariance and geometric primitives in shape recognition. *Cogn. Sci.* 18:325–44
Fawzi A, Frossard P. 2016. Measuring the effect of nuisance variables on classifiers. In *Proceedings of the British Machine Vision Conference (BMVC), York, UK, Sept. 19–22*, art. 137. Durham, UK: Br. Mach. Vis. Assoc.
Felleman DJ, Van Essen DC. 1991. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1:1–47
Field GD, Chichilnisky EJ. 2007. Information processing in the primate retina: circuitry and coding. *Annu. Rev. Neurosci.* 30:1–30
Freedman DJ, Riesenhuber M, Poggio T, Miller EK. 2001. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291:312–16
Gallant JL, Braun J, Van Essen DC. 1993. Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. *Science* 259:100–3
Gallant JL, Connor CE, Lewis J, Rakshit S, Van Essen DC. 1996. Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *J. Neurophysiol.* 76:2718–39
Gallant JL, Shoup RE, Mazer JA. 2000. A human extrastriate area functionally homologous to macaque V4. *Neuron* 27:227–35
Gattass R, Sousa A, Rosa M. 1987. Visual topography of V1 in the Cebus monkey. *J. Comp. Neurol.* 259:529–48
Gattass R, Sousa AP, Gross CG. 1988. Visuotopic organization and extent of V3 and V4 of the macaque. *J. Neurosci.* 8(6):1831–45

Gross CG. 2002. Genealogy of the "grandmother cell." *Neuroscientist* 8:512–18
Gross CG, Rocha-Miranda CE, Bender DB. 1972. Visual properties of neurons in inferotemporal cortex of the macaque. *J. Neurophysiol.* 35:96–111
Halberstadt AG, Saitta MB. 1987. Gender, nonverbal behavior, and perceived dominance: a test of the theory. *J. Personal. Soc. Psychol.* 53:257–72
Hansen KA, Kay KN, Gallant JL. 2007. Topographic organization in and near human visual area V4. *J. Neurosci.* 27:11896–911
Hegdé J, Van Essen DC. 2005. Role of primate visual area V4 in the processing of 3-D shape characteristics defined by disparity. *J. Neurophysiol.* 94:2856–66
Hegdé J, Van Essen DC. 2007. A comparative study of shape representation in macaque visual areas V2 and V4. *Cereb. Cortex* 17:1100–16
Hesse JK, Tsao DY. 2020. The macaque face patch system: a turtle's underbelly for the brain. *Nat. Rev. Neurosci.* 21:695–716
Hinkle DA, Connor CE. 2002. Three-dimensional orientation tuning in macaque area V4. *Nat. Neurosci.* 5:665–70
Hubel DH, Wiesel TN. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160:106–54
Hubel DH, Wiesel TN. 1968. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195:215–43
Hung C-C, Carlson ET, Connor CE. 2012. Medial axis shape coding in macaque inferotemporal cortex. *Neuron* 74:1099–113
Hung CP, Kreiman G, Poggio T, DiCarlo JJ. 2005. Fast readout of object identity from macaque inferior temporal cortex. *Science* 310:863–66
Isik L, Leibo J, Poggio T. 2012. Learning and disrupting invariance in visual recognition with a temporal association rule. *Front. Comput. Neurosci.* 6:37
Ito M, Tamura H, Fujita I, Tanaka K. 1995. Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol.* 73:218–26
Janssen P, Vogels R, Orban GA. 1999. Macaque inferior temporal neurons are selective for disparity-defined three-dimensional shapes. *PNAS* 96:8217–22
Janssen P, Vogels R, Orban GA. 2000a. Selectivity for 3D shape that reveals distinct areas within macaque inferior temporal cortex. *Science* 288:2054–56
Janssen P, Vogels R, Orban GA. 2000b. Three-dimensional shape coding in inferior temporal cortex. *Neuron* 27:385–97
Kobatake E, Tanaka K. 1994. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.* 71:856–67
Koenderink J. 1993. What is a "feature"? *J. Intell. Syst.* 3:49–82
Koenderink JJ, Van Doorn AJ, Kappers AML, Todd JT. 1997. The visual contour in depth. *Percept. Psychophys.* 59:828–38
Kortylewski A, He J, Liu Q, Cosgrove C, Yang C, Yuille AL. 2021. *Compositional generative networks and robustness to perceptible image changes*. Paper presented at the 55th Annual Conference on Information Sciences and Systems (CISS), Baltimore, MD, March 24–26
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* 60:1126–41
Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inform. Proc. Syst.* 25:1097–105
Krumhuber E, Manstead ASR, Kappas A. 2007. Temporal aspects of facial displays in person and expression perception: the effects of smile dynamics, head-tilt, and gender. *J. Nonverbal Behav.* 31:39–56
Kuffler SW. 1952. Neurons in the retina: organization, inhibition and excitation problems. *Cold Spring Harb. Symp. Quant. Biol.* 17:281–92
Li N, DiCarlo JJ. 2008. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science* 321:1502–7

Li SZ, Hou XW, Zhang HJ, Cheng QS. 2001. *Learning spatially localized, parts-based representation*. Paper presented at the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Kauai, HI, Dec. 8–14

Logothetis NK, Pauls J. 1995. Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cereb. Cortex* 5:270–88

Mara M, Appel M. 2015. Effects of lateral head tilt on user perceptions of humanoid and android robots. *Comput. Hum. Behav.* 44:326–34

Marois R, Yi D-J, Chun MM. 2004. The neural fate of consciously perceived and missed events in the attentional blink. *Neuron* 41:465–72

Marr D, Nishihara HK. 1978. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond. B* 200:269–94

Mignault A, Chaudhuri A. 2003. The many faces of a neutral face: head tilt and perception of dominance and emotion. *J. Nonverbal Behav.* 27:111–32

Miller EF. 1962. Counterrolling of the human eyes produced by head tilt with respect to gravity. *Acta Otolaryngol.* 54:479–501

Moeller S, Crapse T, Chang L, Tsao DY. 2017. The effect of face patch microstimulation on perception of faces and objects. *Nat. Neurosci.* 20:743–52

Nandy AS, Sharpee TO, Reynolds JH, Mitchell JF. 2013. The fine structure of shape tuning in area V4. *Neuron* 78:1102–15

Nelissen K, Joly O, Durand J-B, Todd JT, Vanduffel W, Orban GA. 2009. The extraction of depth structure from shading and texture in the macaque brain. *PLOS ONE* 4:e8306

Olshausen BA, Field DJ. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381:607–9

Pasupathy A, Connor CE. 1999. Responses to contour features in macaque area V4. *J. Neurophysiol.* 82:2490–502

Pasupathy A, Connor CE. 2001. Shape representation in area V4: position-specific tuning for boundary conformation. *J. Neurophysiol.* 86:2505–19

Pasupathy A, Connor CE. 2002. Population coding of shape in area V4. *Nat. Neurosci.* 5:1332–38

Pasupathy A, Popovkina DV, Kim T. 2020. Visual functions of primate area V4. *Annu. Rev. Vis. Sci.* 6:363–85

Peterhans E, von der Heydt R. 1989. Mechanisms of contour perception in monkey visual cortex. II. Contours bridging gaps. *J. Neurosci.* 9:1749–63

Poggio GF, Motter BC, Squatrito S, Trotter Y. 1985. Responses of neurons in visual cortex (V1 and V2) of the alert macaque to dynamic random-dot stereograms. *Vis. Res.* 25:397–406

Pospisil DA, Pasupathy A, Bair W. 2018. "Artiphysiology" reveals V4-like shape tuning in a deep network trained for image classification. *eLife* 7:e38242

Pouget A, Dayan P, Zemel RS. 2003. Inference and computation with population codes. *Annu. Rev. Neurosci.* 26:381–410

Pouget A, Deneve S, Duhamel J-R. 2002. A computational perspective on the neural basis of multisensory spatial representations. *Nat. Rev. Neurosci.* 3:741–47

Quian Quiroga R, Kreiman G, Koch C, Fried I. 2008. Sparse but not "grandmother-cell" coding in the medial temporal lobe. *Trends Cogn. Sci.* 12:87–91

Quian Quiroga R, Reddy L, Kreiman G, Koch C, Fried I. 2005. Invariant visual representation by single neurons in the human brain. *Nature* 435:1102–7

Reid RC, Alonso JM. 1995. Specificity of monosynaptic connections from thalamus to visual cortex. *Nature* 378:281–84

Rosenberg A, Cowan NJ, Angelaki DE. 2013. The visual representation of 3D object orientation in parietal cortex. *J. Neurosci.* 33:19352–61

Rosenberg A, Thompson LW, Doudlah R, Chang T-Y. 2023. Neuronal representations supporting three-dimensional vision in nonhuman primates. *Annu. Rev. Vis. Sci.* 9:337–59

Sasikumar D, Emeric E, Stuphorn V, Connor CE. 2018. First-pass processing of value cues in the ventral visual pathway. *Curr. Biol.* 28:538–48.e3

Schworm HD, Ygge J, Pansell T, Lennerstrand G. 2002. Assessment of ocular counterroll during head tilt using binocular video oculography. *Investig. Ophthalmol. Vis. Sci.* 43:662–67

Selfridge O. 1959. Pandemonium: a paradigm for learning. In *Mechanisation of Thought Processes: Proceedings of a Symposium Held at the National Physical Laboratory, November 1958*, pp. 513–26. London: HMSO
Sharpee TO, Kouh M, Reynolds JH. 2013. Trade-off between curvature tuning and position invariance in visual area V4. *PNAS* 110:11618–23
Sheinberg DL, Logothetis NK. 1997. The role of temporal cortical areas in perceptual organization. *PNAS* 94:3408–13
Sigala N, Logothetis NK. 2002. Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature* 415:318–20
Srinath R, Emonds A, Wang Q, Lempel AA, Dunn-Weiss E, et al. 2021. Early emergence of solid shape coding in natural and deep network vision. *Curr. Biol.* 31:51–65
Tanaka K. 1993. *Column structure of inferotemporal cortex: "visual alphabet" or "differential amplifiers"?* Paper presented at the 1993 International Conference on Neural Networks (IJCNN-93), Nagoya, Japan, Oct. 25–29
Tanaka K. 1996. Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19:109–39
Tanaka K, Saito H, Fukada Y, Moriya M. 1991. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J. Neurophysiol.* 66:170–89
Tong F, Nakayama K, Vaughan JT, Kanwisher N. 1998. Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron* 21:753–59
Vaziri S, Carlson ET, Wang Z, Connor CE. 2014. A channel for 3D environmental shape in anterior inferotemporal cortex. *Neuron* 84:55–62
Vaziri S, Connor CE. 2016. Representation of gravity-aligned scene structure in ventral pathway visual cortex. *Curr. Biol.* 26:766–74
Verhoef B-E, Vogels R, Janssen P. 2012. Inferotemporal cortex subserves three-dimensional structure categorization. *Neuron* 73:171–82
Vetter T, Hurlbert A, Poggio T. 1995. View-based models of 3D object recognition: invariance to imaging transformations. *Cereb. Cortex* 5:261–69
Vinje W, Gallant JL. 2000. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287:1273–76
Vogels R. 1999. Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study. *Eur. J. Neurosci.* 11:1239–55
von der Heydt R, Peterhans E. 1989. Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. *J. Neurosci.* 9:1731–48
von der Heydt R, Peterhans E, Baumgartner G. 1984. Illusory contours and cortical neuron responses. *Science* 224:1260–62
von der Heydt R, Zhou H, Friedman HS. 2000. Representation of stereoscopic edges in monkey visual cortex. *Vis. Res.* 40:1955–67
Wang G, Obama S, Yamashita W, Sugihara T, Tanaka K. 2005. Prior experience of rotation is not required for recognizing objects seen from different angles. *Nat. Neurosci.* 8:1568–75
Watanabe H, Pollick FE, Koenderink JJ, Kawato M. 1999. Using motor tasks to quantitatively judge 3-D surface curvatures. *Percept. Psychophys.* 61:1116–39
Watson AB. 2014. A formula for human retinal ganglion cell receptive field density as a function of visual field location. *J. Vis.* 14:15
Wilkinson F, James TW, Wilson HR, Gati JS, Menon RS, Goodale MA. 2000. An fMRI study of the selective activation of human extrastriate form vision areas by radial and concentric gratings. *Curr. Biol.* 10:1455–58
Yamaguchi R, Okamura J, Wang G. 2016. Dynamics of population coding for object views following object discrimination training. *Neuroscience* 330:109–20
Yamane Y, Carlson E, Bowman K, Wang Z, Connor CE. 2008. A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nat. Neurosci.* 11:1352–60
Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS* 111:8619–24
Zhao C, Wang RH, Wang G. 2018. Long-term object discrimination at several viewpoints develops neural substrates of view-invariant object recognition in inferotemporal cortex. *Neuroscience* 392:190–202

Zhou H, Friedman HS, von der Heydt R. 2000. Coding of border ownership in monkey visual cortex. *J. Neurosci.* 20:6594–611

Zhu H, Tang P, Park J, Park S, Yuille A. 2019. *Robustness of object recognition under extreme occlusion in humans and computational models*. Paper presented at the 41st Annual Meeting of the Cognitive Science Society, Montreal, Can., July 24–27

Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank MC, et al. 2021. Unsupervised neural network models of the ventral visual stream. *PNAS* 118:e2014196118

Zikovitz DC, Harris LR. 1999. Head tilt during driving. *Ergonomics* 42:740–46