Responses of neurons in macaque V4 to object and texture images

Justin D. Lieber[1], Timothy D. Oleskiw[1,2], Laura Palmieri[1], Eero P. Simoncelli[1,2], J. Anthony Movshon[1]

[1]Center for Neural Science
New York University
New York, NY 10003

and

[2]Center for Computational Neuroscience
Flatiron Institute
New York, NY 10010

Correspondence: movshon@nyu.edu

Key words: Visual cortex, macaque, object recognition, visual texture, V4

Acknowledgements. We are grateful to Kahlia Gronthos, Sullivan Bacerdo, and Kaitlyn Holman for their assistance, and to Manu Raghavan for his help with hardware and software. This work was supported by grants from the National Institutes of Health (EY022428) and the Simons Foundation (543019) to J.A.M. and E.P.S, as well as the Flatiron Institute of the Simons Foundation (T.D.O. and E.P.S.). J.D.L was supported in part by a Leon Levy Fellowship in Neuroscience. Computational support was provided through NYU IT High Performance Computing services.

Author contributions. Conceptualization and methodology, J.D.L, E.P.S., and J.A.M.; Investigation, J.D.L., T.D.O., and L.P.; Writing, J.D.L., E.P.S., and J.A.M.

Declaration of interests. The authors declare no competing interests.

Abstract

Humans and monkeys can rapidly recognize objects in everyday scenes. While it is known that this ability relies on neural computations in the ventral stream of visual cortex, it is not well understood where this computation first arises. Previous work suggests selectivity for object shape first emerges in area V4. To explore the mechanisms of this selectivity, we generated a continuum of images between “scrambled” textures and photographic images of both natural and man-made environments, using techniques that preserve the local statistics of the original image while discarding information about scene and shape. We measured image responses from single units in area V4 from two awake macaque monkeys. Neuronal populations in V4 could reliably distinguish photographic from scrambled images, could more reliably discriminate between photographic images than between scrambled images, and responded with greater dynamic range to photographic images than scrambled images. Responses to partially scrambled images were more similar to fully scrambled responses than photographic responses, even for perceptually subtle changes. This same pattern emerged when these images were analyzed with an image-computable similarity metric that predicts human judgements of image degradation (DISTS - Deep Image Structure and Texture Similarity). Finally, analysis of response dynamics showed that sensitivity to differences between photographic and scrambled responses grew slowly, peaked 190 ms after response onset, and persisted for hundreds of milliseconds following response offset, suggesting that this signal may arise from recurrent mechanisms.

Introduction

Humans and monkeys are adept at recognizing objects in everyday scenes. The neural substrate for object recognition is a series of computations in the ventral stream of visual cortex (Ungerleider and Mishkin, 1982; Mishkin et al., 1983; Goodale and Milner, 1992; Logothetis and Sheinberg, 1996; DiCarlo et al., 2012; Kaas et al., 2022). This stream consists of a series of hierarchically connected visual areas: beginning in primary visual cortex (V1), continuing through secondary visual cortex (V2) and area V4, before culminating in inferotemporal cortex (IT). Population responses in IT can successfully discriminate visual objects (Pasupathy and Connor, 2002; Rust and DiCarlo, 2010) and predict human performance on object recognition tasks (Majaj et al., 2015). However, the intermediate computations that build these object selective responses are not yet well understood.

One influential hypothesis proposes that early visual areas use parallel channels to measure different visual substances (“stuff”) - e.g. vertical or horizontal lines, redness or blueness, or different directions of moving energy (Adelson and Bergen, 1991). These early channels are processed in turn by later visual areas, resulting in new channels that are selective for elaborated conjunctions of the original feature set (Fukushima, 1980; Riesenhuber and Poggio, 1999; DiCarlo and Cox, 2007; Yamins et al., 2014). This framework suggests that neurons in mid-level visual areas like V2 and V4 will exhibit selectivity for the complex, textural features present in the natural world.

Consistent with this hypothesis, neurons in area V2 have been found to be selective for the statistical regularities of natural texture images, unlike neurons in area V1 (Freeman et al., 2013; Ziemba et al., 2016, 2018, 2019). Specifically, image statistics derived from a V1-like representation (Portilla and Simoncelli, 2000) were used to generate matched sets of texture images containing either “naturalistic” structure or spectrally matched noise. Neurons in V2 responded more strongly to naturalistic images than to matched noise (Freeman et al., 2013), and this modulation emerged slowly over the course of 60-80 ms following response onset (Freeman et al., 2013; Okazawa et al., 2016; Ziemba et al., 2018). Compared to V2, neurons in V4 show stronger modulation by naturalistic texture that takes even longer to emerge (Okazawa et al., 2016; Lee et al., 2024, 2025).

An alternative perspective on ventral stream computation has focused on object-centric representations (“things”) that first emerge in V4 (Pasupathy et al., 2020). Many neurons in V4 respond robustly and selectively to object curvature and shape (Pasupathy and Connor, 1999, 2001), with responses that are invariant to object position (Pasupathy and Connor, 1999; Rust and DiCarlo, 2010; Nandy et al., 2013; Sharpee et al., 2013; El-Shamayleh and Pasupathy, 2016) and that integrate global context about border ownership (Bushnell et al., 2011). Results from neurophysiology

(Rust and DiCarlo, 2010; Kramer et al., 2023) and human fMRI (Movshon and Simoncelli, 2014; Long et al., 2018) suggest that V4 is more strongly modulated by photographic images than by matched Portilla-Simoncelli textures.

To parse the interplay between “stuff” and “things,” we can define different levels of image feature complexity. We define simple spectral statistics, which are sufficient to account for the selectivities of neurons in V1, as “low-level” features. Coactivations of these spectral features provide the basis for the Portilla-Simoncelli texture synthesis model, and are the features that drive V2’s selectivity for naturalistic texture. We define these coactivations as “mid-level” features. We define “high-level” features as anything not captured by mid-level coactivations. When high-level features are removed from natural images, the characteristic features that distinguish objects from textures are destroyed (Freeman and Simoncelli, 2011; Long et al., 2018). It is not well understood how neuronal selectivity in V4 depends on these different kinds of features.

In this study, we examined responses of macaque V4 neurons to a continuum of images that span the space between “things” (photographic images) and “stuff” (scrambled textures). Using an adaptation of the Portilla-Simoncelli texture synthesis algorithm (Freeman and Simoncelli, 2011), we synthesized images that matched the complex features of an original photograph in spatially localized regions. By varying the region sizes from small to large, we generated images that smoothly transition between photographic images and scrambled textures. We recorded the responses of single neurons in V4 to these images, and found that neural populations could reliably detect the presence of photographic image features that were destroyed by image scrambling. V4 populations could reliably discriminate between photographic images, and were more sensitive when discriminating between photographic images than between scrambled images. While photographic images and scrambled images drove similar response intensity in V4 neurons, photographic images drove a larger dynamic range of response than scrambled images, supporting stronger image discrimination. Photographic image responses were highly sensitive to even small amounts of image scrambling, and this sensitivity could be predicted by DISTS, an image quality assessment metric that was built to capture perceptual sensitivity to image distortions. Finally, V4’s sensitivity to photographic structure emerged slowly relative to the onset of neural activity, and persisted well after stimulus onset, suggesting that it arises from ongoing recurrent or feedback processes.

Methods

Image generation

We chose a core set of 20 large photographic images from two databases. Half were taken at a baboon habitat in Botswana (the “UPenn Natural Image” database, Tkačik et

al., 2011), intended as an approximation of the evolutionary context for primate vision. The other half feature everyday objects in their appropriate context (the “Reachspace” database, Josephs et al., 2021). These images of objects on natural backgrounds are similar to those typically used to train and evaluate deep neural network models of area V4 (Yamins et al., 2014). All source images were cropped and resolution adjusted so as to contain objects centered within a 1280x1280 pixel frame. From each source image, we then cropped four overlapping 512x512 subimages, centered at locations +/-56 pixels horizontally and vertically displaced from the center of the source image. In the experiment, images were presented on a screen at a resolution of 80 pixels per degree. Therefore, when presented to the subject, these “shifts” of the underlying parent image correspond to an angular distance of 1.4 deg.

To generate scrambled images, we used an adaptation of the Portilla & Simoncelli texture synthesis model that measures and synthesizes images with statistics that are matched within localized pooling subregions (Freeman and Simoncelli, 2011, https://github.com/freeman-lab/metamers). For this study, we arranged these pooling regions as a square grid of smoothly overlapping fields that tiled the image. Within each region, we measured a set of texture statistics (Portilla and Simoncelli, 2000). Specifically, we processed the image by convolving it with a bank of 32 oriented filters (4 orientations, 4 scales, 2 phases), computed with periodic boundary conditions. To create both “simple-cell-like” and “complex-cell-like” channels, we computed both the linear and amplitude responses of each filter. To create a set of mid-level statistics, we computed the pairwise product of these “simple” and “complex” responses at different positions, orientations, and scales. Finally, we computed weighted averages of these products within each pooling region, resulting in a set of covariance statistics.

We synthesized new “scrambled” images by initializing an image with Gaussian white noise, then iteratively adjusting the pixels to match the measured texture statistics within each subregion. To prevent boundary artifacts from appearing within the synthesized images, we defined a circular vignette with diameter 492 pixels (6.15 deg) and raised cosine edges spanning 25 pixels (0.31 deg). Between iterations of the synthesis we blended the edge regions of the image (using blending fractions governed by the vignetting function) with the pixel values of the original photographic image. After synthesis was complete, the images were cropped with this same circular vignette to remove these “photographic” regions, such that only “scrambled” portions of the image remained.

For each of the 4 “shifted” subimages, we synthesized new images based on 1x1, 2x2, 3x3, 4x4, and 6x6 grids of statistical pooling windows. When presented to the subject, these pooling regions of these grids subtended 6.4, 3.2, 2.1, 1.6, and 1.1 deg, respectively. We also included the original (unscrambled) images, giving a total of 6

distinct scrambling conditions. We refer to the set of 24 images derived from an original photograph (4 shifts x 6 pooling window conditions) as an “image family.”

Experimental procedures

Experimental procedures for monkeys conformed to the National Institute of Health *Guide for the Care and Use of Laboratory Animals*, and were approved by the New York University Animal Welfare Committee.

We recorded eye position with a high-speed, high-precision eye tracking system (EyeLink 1000). The animal initiated each trial by fixating on a small dot (~0.25 degrees wide), and maintained fixation within a window of 1-2 degrees. Images appeared for 200 ms, with a 200 ms inter-stimulus interval, and were blocked in 3-8 consecutive presentations, after which the animal received a juice reward.

Acute recordings

Under general anesthesia, we implanted one male animal (Monkey M) with a titanium head post and recording chamber over area V4. We recorded single unit activity within V4 using both single site electrodes (FHC) and a linear microelectrode array (Plexon S-probe, 64 channels with 50 μm spacing, over 3.2 mm total). We identified the location of the lunate sulcus, then recorded in surface V4 anterior to the sulcus.

For single-site electrode recordings, we recorded all cells we found that were well-isolated and could be reliably driven by the image set. For linear array recordings, we inserted the probe into cortex deeply enough that visual stimuli drove multi-unit responses on most channels. The centers of neuronal receptive fields lay between 1 and 12 degrees from the center of gaze (median=6.1 deg). We isolated single units from multi-site recordings using KiloSort 2.5 (Steinmetz et al., 2021) followed by manual curation of well-isolated spikes (Phy, https://phy.readthedocs.io/). Cells were only analyzed when recordings lasted long enough to get at least 4 repetitions to all image stimuli (50/82 for single-site electrodes, 117/123 linear array), and when the evoked firing rate to all image stimuli was at least 2 spikes/second (44/50 single-site electrodes, 79/117 linear array), for a total of 123 analyzed units from Monkey M.

Chronic array recordings

Under general anesthesia, we implanted a 96 electrode “Utah” recording array (Blackrock) in area V4 of one male animal (Monkey B). We identified area V4 as the gyrus located between the lunate and superior temporal sulci, and dorsal to the tip of the inferior occipital sulcus. Electrodes shanks were 1 mm long and were arranged in a square grid (spacing 0.4 mm).

We recorded neural responses to our stimulus set over a period of five weeks, during which we collected 15 recording sessions. All recorded neurons had receptive field centers located between 1 and 3 degrees of eccentricity. We isolated single-units from these recordings using Mountainsort 5 (Chung et al., 2017) with SpikeInterface (Buccino et al., 2020), followed by manual curation of well-isolated spikes (as above). Because recording sessions on an electrode from consecutive days could potentially contain the same units (Dickey et al., 2009), we chose to use a subset of the recording sessions for full analysis. Specifically, for each electrode, we characterized the ability of every recorded single unit to discriminate image responses from blank trial responses. We then selected the single day of recording that had the highest level of stimulus vs. blank discriminability. After removing electrode channels that never showed responsive single units, this resulted in a final population of 88 single-unit recordings from Monkey B.

Firing rate analyses: For all cells, we computed an average firing rate on each trial in an interval between 50 ms and 400 ms after stimulus onset. When comparing or combining data across neural populations, we computed a normalized firing rate by dividing all individual trial firing rates by the average firing rate response to all non-blank stimulus conditions.

Stimulus-dependent fractional variance (SDFV): The visually evoked firing rate responses of V4 neurons varied across both repeated presentations of the same stimulus, as well as different presentations of different stimuli. We sought to quantify the extent to which differences in stimulus responses were reliable. To this end, we computed the SDFV as the proportion of overall variance that could be explained by differences in stimuli. Specifically, we computed total variance as the variance in firing rate responses across all trials, and we computed stimulus-dependent variance as the variance across each stimulus's trial-averaged firing rate. SDFV was computed as the ratio of stimulus-dependent variance to total variance.

Modulation index: We computed a modulation index for individual cells as a standardized measure of the signed strength of photographic vs. scrambled firing rates.

In Figure 2A, we computed a modulation index of individual neurons to individual image families. We did this by averaging firing rates across image shifts, to find an average firing rate for photographic images and an average rate for scrambled images. We then computed this modulation index as the difference between these rates, divided by their sum.

We also computed a modulation index for each neuron. We first averaged firing rates over all image family and shift conditions to find an average rate for both the photographic condition and the fully scrambled condition. We again computed the modulation index as their difference divided by their sum.

Photographic-scrambled correlation: For each neuron, we computed the correlation across families between its response to photographic images and its response to scrambled images. To compute a family's evoked firing rate, we averaged firing rates over all trials and all shifts. We then computed the Pearson correlation between average photographic rates and average scrambled rates.

Rate-modulation correlation: For each neuron, we quantified whether the image families that most strongly drove the cell were also the image families that were most strongly modulated by photographic structure. To compute an image family's evoked firing rate, we averaged firing rates over all shifts and over both photographic and fully scrambled conditions. To compute each image family's modulation, we averaged firing rates over all shifts, then computed the difference between the photographic and scrambled firing rates. We then computed each neuron's Pearson correlation between family-evoked rates and family-evoked modulation.

Image family ranking: We sought to visualize whether the image families that most strongly drove neurons were also the families most strongly modulated by photographic structure. Our first thought was to simply rank each neuron's preferred image families, but we were concerned that trial-to-trial fluctuations in firing rate might lead to overfitting that obscured the true magnitude of photographic image modulation. We computed a cross-validated ranking of image families by splitting the image shift conditions into training sets (2 shift conditions) and testing sets (2 shift conditions). As with rate-modulation correlation, we averaged firing rates over shift conditions and over photographic and fully scrambled conditions, then used these rates to rank image families. We then recomputed average firing rates for the photographic and scrambled conditions from the test set data, and sorted those values using the training set rankings. The rank order plots in Figure 3 are the average of two rank order plots made from the two possible partitions of training/testing data.

Classification with LDA: For a range of different tasks, we sought to quantify the ability of neural populations to discriminate between pairs of classes of stimuli. To this end, we implemented a cross-validated linear discriminant analysis (LDA). We first split our image set in half such that, for each image family, two diagonally opposite shift locations were used for model training, and the remaining two were used for testing.

For population classification tasks, we split our population into two matched sets of 105 neurons. We repeated this resampling process over 100 distinct splits of the full population into two distinct subpopulations. For each of these subpopulations, for every we constructed pseudo-trials by randomly selecting a single trial from each neuron. We repeated this process to create 500 pseudo-trials (with replacement) for each image.

We then used LDA to learn the weighted sum across neurons that best separated the two classes:

$$W = (\, cov(x_1) + cov(x_2)\,)^{-1} * (\bar{x}_2 - \bar{x}_1)$$

where x_1 and x_2 are matrices of training set firing rates corresponding to the two classes, respectively, with columns of neurons and rows of pseudo-trials of all images in the class, and W is the vector of resulting weights on the neurons. We then apply these weights to the testing data to compute a single scalar value for each trial, and then compute discriminability on these scalar values:

$$d' = \frac{\bar{w}_2 - \bar{w}_1}{\sqrt{0.5*(var(w_2)+var(w_1))}}$$

where w_1 and w_2 are the weighted sums over testing set data for the two classes, respectively.

We trained classifiers for four distinct discrimination tasks:

1) Stimulus vs. blank: We trained classifiers to separate 1) a combined set of all image conditions (all image families, all scrambling conditions, and two shifts per training/testing set) and 2) blank trials. Blank trials were split in half into training and testing sets.

2) Family-specific scramble vs. structured: For each image family, we trained a classifier to separate fully scrambled images from either partially scrambled conditions or from the original photographic images. We averaged performance across all image families to get final discriminability values.

3) Family-general scramble vs. structured: For family-general discrimination, we trained a single discriminant, shared across families, between 1) the full set of scrambled images and 2) similar combined sets of either partially or photographic image conditions.

4) Pairwise family: For each scrambling condition, we identified every possible pair of two image families. We trained separate discriminants for every pairwise condition. To get a final discriminability for each scrambling condition, we averaged performance over all pairs of families. Additionally, we sought to quantify how well pairwise family classifiers generalized between photographic images and fully scrambled images. For this generalization condition, we took the discrimination axes learned from fully scrambled image training set responses, and applied them to the corresponding photographic testing set responses.

In the standard analysis for all four tasks, sensitivities were computed using average firing rates computed between 50 and 400 ms following stimulus onset. For the dynamics analysis, discriminabilities were computed in a series of 50 ms bins, spaced every 5 ms. For both standard and dynamics analyses, we computed final discriminabilities by averaging over both permutations of the trial splits into training and testing image sets, over both disjoint neural populations within each population resampling, and over all population resamplings. To compute error bars, we measured sample variances across the disjoint pair of neural populations, then averaged sample variances over the resampled populations.

Perceptual and neural distance metrics

We sought to evaluate whether the patterns of response we observed in V4 could be explained by computational models of perceptual image similarity that are used to quantify the perceived difference between an original image and a corrupted counterpart (e.g. to measure the quality of a lossy file format). We computed pairwise distances for both the V4 population response and for a set of image similarity metrics. In all cases, we only computed distances within sets of images of the same family and the same shift (an original photographic image and its partially or fully scrambled counterparts). Each of these sets had 6 images, which resulted in 30 total comparisons per “shifted” photographic image. Over 20 image families and 4 shifts per family, this resulted in 2,400 total distances.

Neural distance: We first split our neural population into two matched subpopulations of 105 neurons. We repeated this resampling process over 100 distinct splits. Within each subpopulation, for each image we found a 105-neuron vector of averaged, normalized firing rates. We then computed the Euclidean distance of each pair of vectors, and normalized that distance by the number of neurons.

Image similarity metrics: We used 3 standard image similarity metrics to compare different images: RMS pixel distance, structural similarity (SSIM - Wang et al., 2004), and Deep Image Structure and Texture Similarity (DISTS - Ding et al., 2022). For RMS pixel distance, we first normalized each image so that 0 and 1 corresponded to minimum and maximum luminance. We then computed the RMS difference over pixels for each pair of images. We computed SSIM using the *ssim()* function provided in Matlab, with all exponents set to 1. For the DISTS metric, we used a Matlab implementation of the algorithm provided by the authors (https://github.com/dingkeyan93/DISTS).

As a summary statistic, we computed the Pearson correlation between each image similarity metric and the corresponding neural distance. Correlations were averaged over all subpopulations. To compute error bars, we measured the sample variance

across each disjoint pair of neural populations, then averaged sample variances over the resampled splits of the population.

Results

Creating a continuum from scrambled textures to photographic images

To create an image set that spanned the continuum between objects and textures, we first selected 20 photographic images from two natural image databases. Half were taken from the “Birthplace of the Eye” image set, containing evolutionarily relevant photographs taken at a primate reserve in Botswana (Tkačik et al., 2011), and the other half from the “Reachspace” image set (Josephs et al., 2021), containing scenes of manmade objects in their typical context. From each of these images we extracted 4 discrete image “shifts” (Figure 1A), which contain similar (but not identical) content. We then applied varying “levels” of scrambling to these images, by parametrically varying the size of the local pooling regions over which we measured texture statistics and synthesized new images (Figure 1B). Images synthesized with smaller pooling regions were more similar to their progenitor photographs than those synthesized with larger pooling regions. Six scrambling levels were chosen to cover the range from the original photographs to their globally scrambled counterparts, in approximately equal perceptual increments (Figure 1C). In total, counting all shifts and pooling region sizes, from each original photograph we created an image “family” containing 24 images.

We recorded the responses of 211 single unit recordings in area V4 presentation of these stimuli over their receptive fields (see Methods for selection criteria). Nearly all neurons responded significantly more to the image set than to a blank screen (197/211 cells significantly modulated, $p<0.05$, permutation test). To quantify how much V4 neuronal responses were modulated by image identity, we computed the stimulus-dependent fractional variance (SDFV), the proportion of response variance that was captured by stimulus-driven modulation, rather than trial-to-trial noise (see Methods). Because SDFVs were typically greater than permuted controls (median SDFV 0.369±0.180 median absolute deviation (MAD), 193/211 neurons with $p<0.05$, permutation test), we concluded that V4 neurons were tuned to differences between the images.

The primary goal of this study was to determine whether V4 tuning was modulated by the presence of objects in photographic images. We examined both object detection (whether V4 could discriminate between photographic images and their fully scrambled counterparts) and discrimination (whether V4 could distinguish objects, and how that capability was degraded by scrambling). These two approaches can be visualized in Figure 1C as discriminations across either rows (for the photograph vs. scramble task) or columns (for the family discrimination task), respectively.

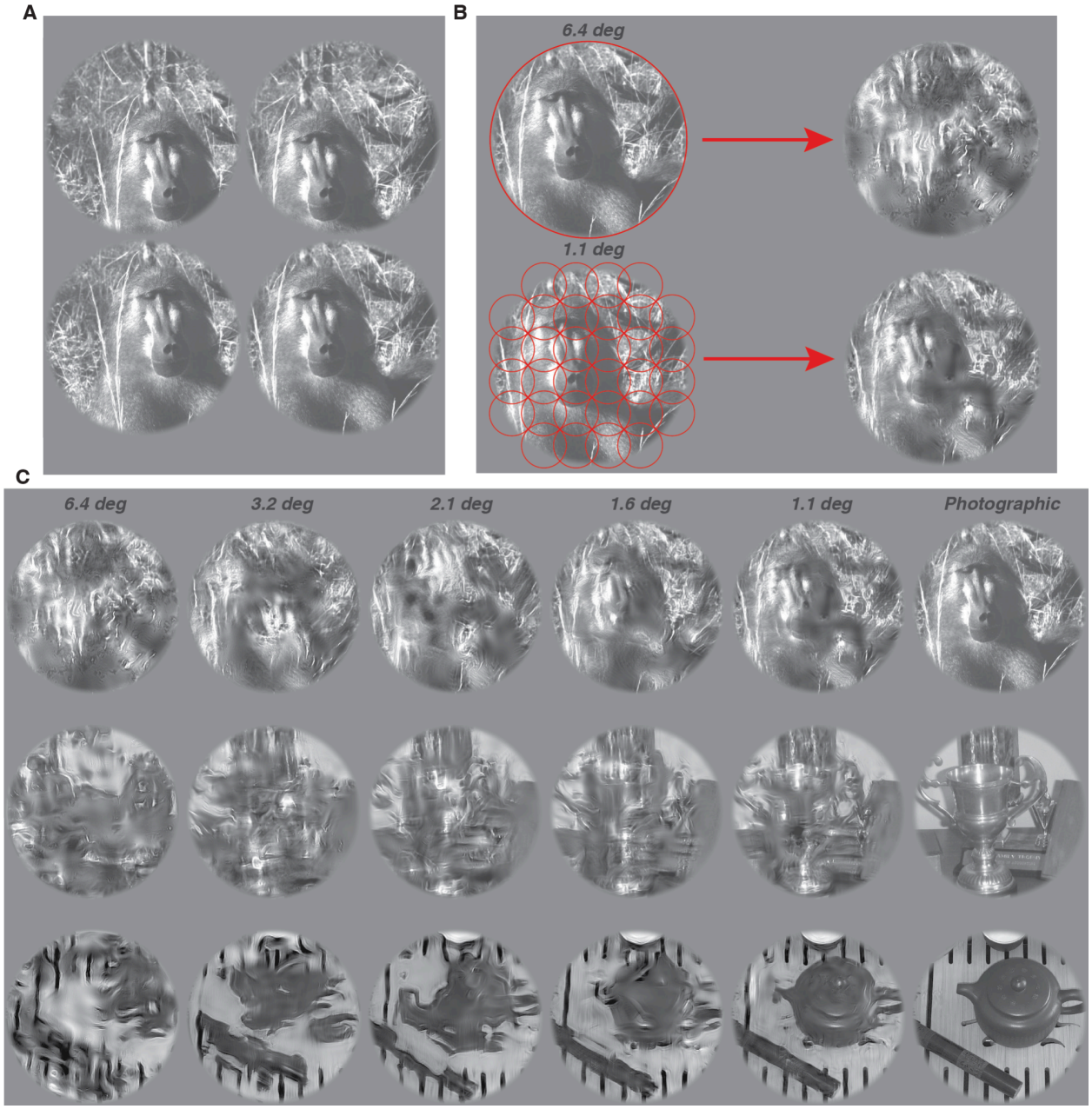

Figure 1. Creating a continuum from scrambled textures to photographic images. A) We cropped four distinct shifted subimages from each source image. These subimages were overlapping, and contained the same central features. B) We used the Portilla-Simoncelli texture model to scramble image features within local pooling regions (red circles). When a single pooling region covered the image (top) the content of the image was fully scrambled. When multiple smaller pooling regions covered the image (bottom), the local features within each pooling region were scrambled but the image retained some of the global structure of the original. C) By varying the size of the pooling regions used in the image synthesis algorithm, we created a continuum of images between scrambled textures and the original images. We used 5 pooling region sizes across our 6.4 degree diameter image: 6.4, 3.2, 2.1, 1.6, and 1.1 degrees.

Images smoothly transitioned from scrambled textures (left) to original photographic images of objects (right).

V4 responses can detect photographic structure

We evaluated whether V4 responses could distinguish between photographic and scrambled images. Because these images were matched in their texture statistics, this discrimination relied on detecting high-level, object-like features in the photographic images. Therefore, this discrimination can be thought of as a difficult object detection task - the ability to separate “things” from “stuff” when low- and mid-level features of the images are matched.

We quantified how strongly V4 responses were modulated by object-like structure by defining a modulation index (MI): the difference between firing rate responses to photographic and scrambled images, divided by their sum. Modulation varied across neurons and across image families, from strongly positive (photograph-preferring) to strongly negative (scramble-preferring) (Figure 2A). On average, the V4 neurons did not show a preference for photographic images over scrambled images (median MI for individual neurons = -0.002±0.050 MAD) (Figure 2B). A substantial, and similar, number of individual neurons exhibited either significantly positive (56/211, $p<0.05$, permutation test) or significantly negative (59/211) modulation.

Both positively modulated and negatively modulated neuronal responses are informative about photographic structure. We quantified this informativeness by measuring the sensitivity of classifiers trained to discriminate photographic images from their scrambled counterparts. For each image family, we trained a classifier on the responses to two distinct shifts of the photographic and scrambled images. To measure discriminability, we applied that classifier to the responses to the other two image shifts. This cross-validation procedure ensured that our measured sensitivities were invariant to small shifts in the underlying image, and thus that they relied on general properties of the image family instead of simple, pixel-specific features.

We first measured discriminability for individual V4 neurons. For each image family, we quantified discriminability as a cross-validated d-prime. Then, for each neuron, we averaged d-primes across all families. Most neurons in the population could reliably discriminate between photographic images from scrambled images (Figure 2C). We also trained classifiers to discriminate between images with intermediate levels of scrambling and those images’ scrambled counterparts (Figure 2D, white points). Detection of object structure was much weaker for partially scrambled images.

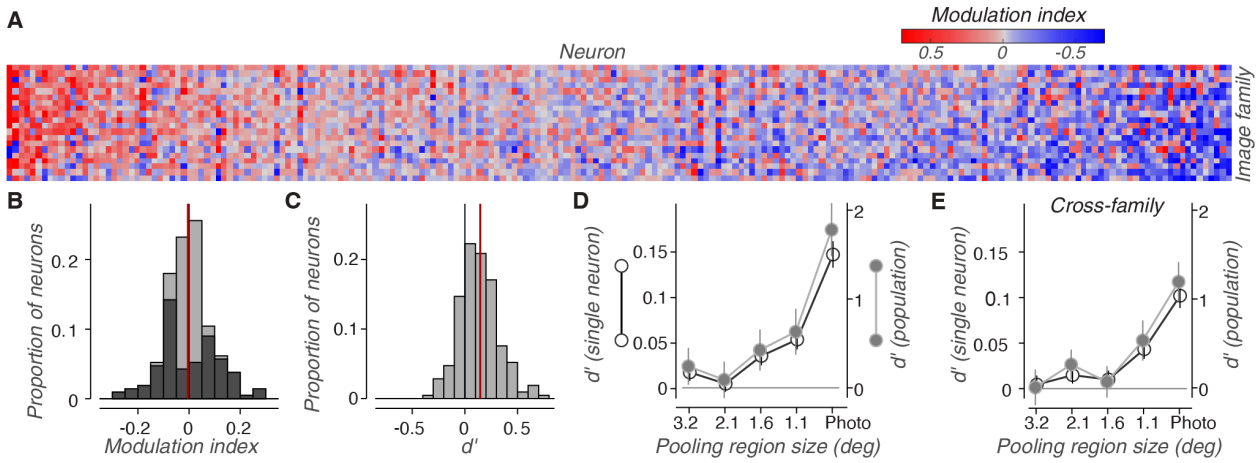

Figure 2: V4 reliably discriminates photographic scenes from their scrambled counterparts. A) Modulation indices of 211 neurons computed separately for each of 20 image families. Both families and neurons are ordered from highest to lowest average modulation index (top to bottom, and left to right, respectively). B) Distribution of the modulation indices of individual neurons, calculated on responses averaged across all image families. Neurons with modulation indices that are significantly different from 0 ($p<0.05$) are labeled in dark grey. Nearly as many neurons were significantly positively modulated (N=56, $p<0.025$) as significantly negatively modulated (N=59, $p<0.025$). The red line denotes the population average (MI=-0.004). C) Distribution of neuronal discriminabilities between paired photographic and scrambled images, measured as cross-validated d'. Discriminabilities are averaged across image families. Negative values of d' correspond to mismatches between testing and training data – that is, unreliable tuning. The red line denotes the population average (d'=0.14). D) Discriminability between fully scrambled image responses and either the original photographic images, or intermediate levels of scrambling, measured for both individual neurons (black) and neural populations (grey, population size=105 neurons). Discriminabilities were computed individually for each image family, then averaged across families. Compared to the single neuron axis, the population axis is scaled by the square root of the population size (~10.2), which is the prediction corresponding to a population of N independent detectors. Error bars show either the standard error across neurons or the standard deviation across neural populations, respectively. E) Discriminability between fully scrambled image responses and either the original photographic images, or intermediate levels of scrambling, now for classifiers trained on all image families simultaneously. All plotting conventions as in D)

Next, we measured the discriminability of population responses in V4. Using a cross-validated linear discriminant analysis (LDA), we found the weighted sum of neuronal responses that best separated photographic from scrambled images, then quantified discriminability by computing d-prime values along the resulting discriminant axes (see *Methods: Classification with LDA*). Small populations of neurons (Figure 2D, population size=105 neurons) could reliably discriminate photographic from scrambled images, and performed more poorly for partially scrambled images (64.5% lower for the least scrambled condition, 1.1 deg, compared to photographic images). That is, adding

a small amount of scrambling to the photographic images led to a large change in discriminability. If the neurons in these populations were encoding information independently, discriminability should scale with the square root of population size (Green and Swets, 1966). We found this to be the case (Figure 2D, compare white points to gray points).

These results illustrate that V4 responses are *selective* for the presence of objects. Next, we investigated whether this selectivity was *invariant* to changes in the specific objects present in the image. In other words, we asked whether V4 contains a unified representation of object structure: a single readout that could support the perceptual ability to distinguish photographic images from their scrambled counterparts (Wallis et al., 2019). To this end, we retrained our photographic vs. scrambled classifiers using the full, combined set of image families (Figure 2E). These cross-family classifiers could also reliably distinguish photographic and scrambled images, with only a 32.4% reduction in sensitivity compared to the average performance of family-specific classifiers (Figure 2D). Like family-specific classifiers, family-general classifiers were more sensitive to photographic images than partially scrambled images (54.9% lower for 1.1 deg scrambled). In summary, V4's sensitivity to object structure is largely invariant to changes in the specific objects within the scene.

V4 responses discriminate photographic images better than scrambled images

Next, we evaluated whether V4 responses could distinguish pairs of photographic image families. This is a type of object recognition task, which is posited to be a primary objective of ventral stream processing (DiCarlo et al., 2012). As before, we quantified this ability by training classifiers that were cross-validated across different shifts within each image family. In other words, these classifiers measured the robustness of object recognition to small translations of the image ("position invariance", as in Rust and DiCarlo, 2010; El-Shamayleh and Pasupathy, 2016).

We first measured image family discriminability for individual V4 neurons. For each pair of image families, we quantified discriminability as a cross-validated d-prime. We then averaged d-primes over all pairs of families. Single-cell discriminability was typically reliable for both photographic image responses and for scrambled image responses (Figure 3A), and was systematically lower for scrambled than photographic image responses (33.9% lower). Responses to intermediate levels of scrambling produced intermediate levels of sensitivity (Figure 3B, white points).

Next, we measured discriminability for V4 population responses. As before, we used LDA to find decoding axes that best separated pairs of image families, then measured discriminabilities along those axes (see *Methods: Classification with LDA*). Previous work has shown that populations of V4 neurons can better classify images of

photographs than images of matched textures (Rust and DiCarlo, 2010). Consistent with those results, and as shown in Figure 3B (gray points), we found that population discriminability for scrambled image responses was ~2/3 as large as for photographic responses. Intermediate levels of scrambling resulted in intermediate measurements of discriminability. However, as in the photographic detection task, we observed that adding a small amount of scrambling to photographic images led to a large change in discriminability. The gap between photographic images and their closest, least scrambled counterparts (1.1 deg. scrambling regions) was as large as the gap between those least scrambled images and the fully scrambled images ($\Delta d'=0.619$ vs. $\Delta d'=0.616$, respectively).

The texture synthesis procedure generates scrambled images containing low- and mid-level features that are matched to their corresponding photographic images, and that only differ in their high-level features. The extent to which V4 neurons are tuned to these low- and mid-level features should be reflected by similar responses to matched photographic and scrambled images, and similar discriminabilities. To quantify tuning similarity for individual neurons, we measured the correlation between photographic and scrambled image responses. Consistent with previous reports (Long et al., 2018; Kramer et al., 2023), we found that for most neurons, photographic image responses were well correlated with scrambled image responses (Figure 3C, median correlation = 0.443±0.21 MAD across all neurons). This suggests that low- and mid-level features explain a meaningful proportion of V4 tuning.

This prompted us to ask how strongly low- and mid-level features contributed to image family discriminability. We reasoned that a classifier trained on scrambled image responses would only be aware of low- and mid-level features, and would not be able to take advantage of the characteristic high-level features that distinguish photographic images. With this in mind, we measured how well classifiers trained on scrambled image responses generalized when tested on photographic image responses. We found that these classifiers performed with similar discriminability when tested with either scrambled responses or with photographic responses (Figure 3B, “Gen.”) — that is, ~2/3 as large as for classifiers that were trained and tested on photographic responses. This suggests that scrambled images drive patterns of activity in V4 that are entirely preserved when responding to photographic images. It also suggests that high-level features in photographic images drive patterns of activity that are independent from those driven by low- and mid-level features, and that these new patterns fully explain the ~1/3 difference in image family discrimination between photographic and scrambled responses.

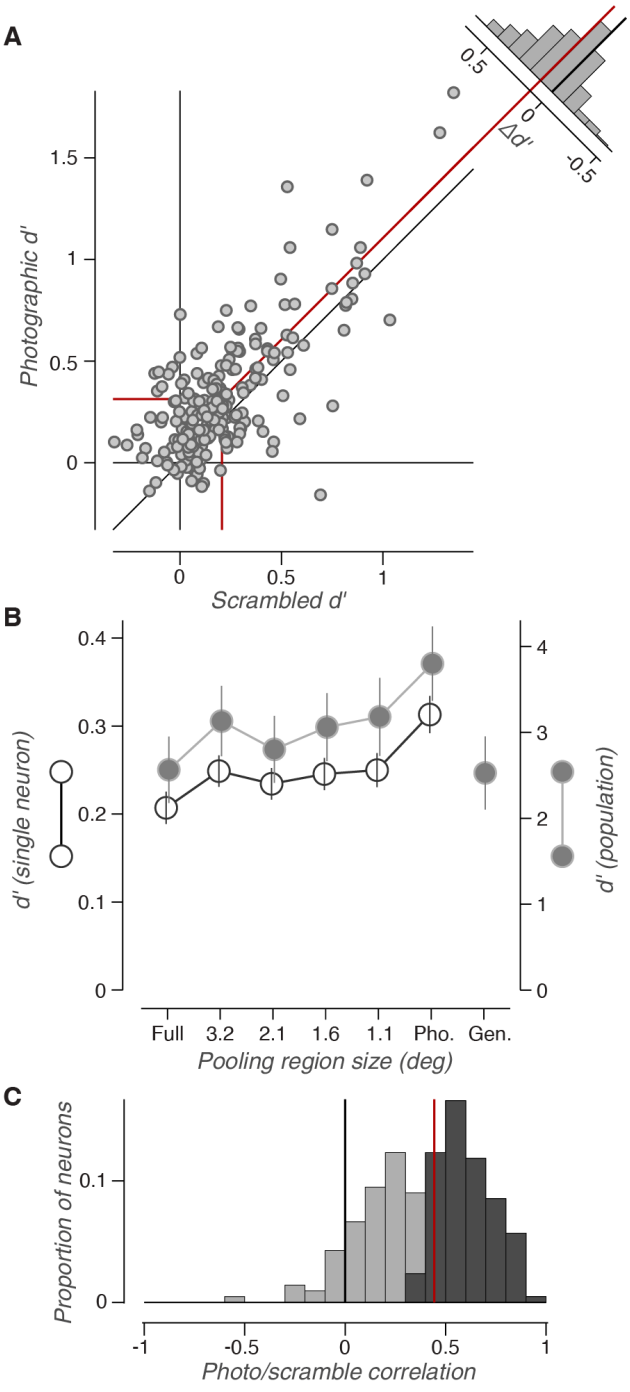

Figure 3: V4 discriminates photographic scenes more reliably than their scrambled counterparts. A) Neuronal family discriminability (d') for photographic images, plotted against family discriminability for scrambled texture images. Discriminability was measured as a cross-validated d' measure, averaged across all possible pairs of image families. The inset above shows the neuronal distribution of the difference between photographic and scrambled d'. Red lines denote mean values for the photographic distribution (d'=0.23), the scrambled distribution (d'=0.15), and their difference (Δd'=0.10). A majority of neurons (150/211) show higher discriminability between photographic than scrambled images. B) Average family discriminability (d') for full and intermediate levels of scrambling is shown for both single neurons (black) and neural populations (grey, population=105 neurons). As in Figure 2D-E, the population axis is scaled relative to the single neuron d' axis by the square root of the population size (~10.2), which is the corresponding prediction for a population of N independent detectors. Error bars show either the standard error across neurons or the standard deviation across neural populations, respectively. The far right point ("Gen.") shows the population discriminability of a classifier trained on scrambled images and tested on photographic images. Both the "Full" and "Gen." conditions are ~2/3 the discriminability of the photographic condition. C) Distribution of individual neuron correlations between firing rate responses to photographic images and to matched scrambled images. Neurons with correlations significantly greater than 0 are labeled in dark grey (122/211, $p<0.05$). The red line denotes the median ($r=0.443$).

Neurons respond with greater dynamic range to photographs than to scrambled textures

We next asked how individual neurons were tuned to the high-level features present in different photographic image families. Based on the conjecture that stronger responses might be associated with stronger effects, we predicted that the image families that most strongly drove an individual cell might also be the ones that were most strongly modulated by photographic structure. We defined an image family's average drive to a neuron as the average response to all photographic and fully scrambled images in that family. For most neurons, the average drive per family was positively correlated with the difference in response between photographs and scrambled images for that family (Figure 4A, median correlation = 0.21±0.23 MAD, $p<0.001$, permutation test). That is, the image families that most strongly drove a neuron also tended to be the families that exhibited the largest response difference between photographic and scrambled images.

To summarize these effects, we developed a cross-validated ranking procedure to determine which image families evoked the strongest responses for each neuron (see *Methods: Image family ranking*). When we averaged the results of this procedure across the full population, we found that the families that evoked the strongest responses were positively modulated, and the families driving the weakest responses were negatively modulated (Figure 4B).

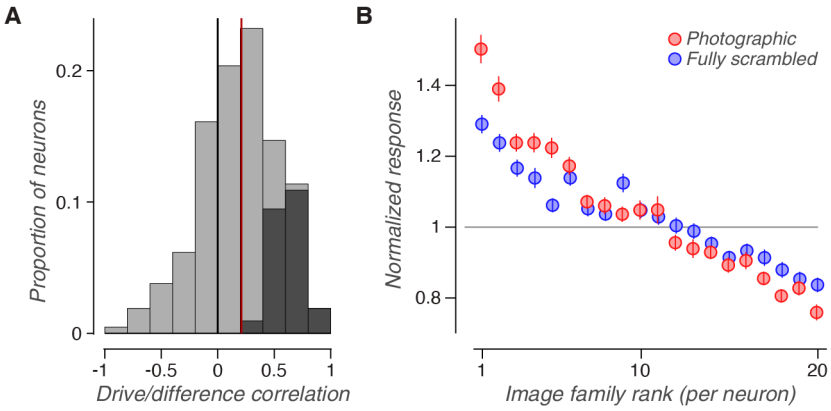

Figure 4: Neurons respond with greater dynamic range to photographs than to scrambled textures. A) Distribution of individual neuron correlations between the average drive to an image family (averaged over both photographic and scrambled images) and that family's scrambling difference (the difference between photographic and scrambled responses). Neurons with correlations significantly greater than 0 are labeled in dark grey (49/211, $p<0.05$). The red line denotes the median value ($r=0.209$). This positive value indicates that the image families that most strongly drove a neuron also tended to be positively modulated. B) Normalized firing rates, averaged over all neurons, against image family rank. Rankings were calculated, separately for each neuron, as a cross-validated measurement of each image family's average drive. Error bars denote standard errors across neurons and shifts. Strongly driven families are positively modulated and weakly driven families are negatively modulated.

The per-cell ranking procedure in Figure 4B illustrates that, on average, photographic images drove cells to both their highest firing rates (for strongly driving families) and their weakest firing rates (for weakly driving families). That is, photographic images drove a wider dynamic range of responses than scrambled images, as suggested by previous reports (Rust and DiCarlo, 2010). We quantified this effect using the variance of normalized firing rates, which was larger for photographic images (median σ^2=0.098) than scrambled images (median σ^2=0.082). Moreover, 152 of 211 neurons (72%) had a larger response variance for photographic images than for scrambled images. This wider dynamic range may explain the superior classification performance of the V4 population for photographic images over scrambled images.

A model of perceptual similarity predicts V4 responses to image scrambling

For both photographic structure detection and image family discrimination, we observed that partially scrambled images drove very different responses when compared to photographic images. This was surprising to us, as we chose our scrambling levels to roughly correspond to equal perceptual increments (Figure 1C). By inspection, examples from the least scrambled condition (pooling size 1.1 deg, see Figure 1C) appear quite similar to the original photographs. Nonetheless, classification performance differed strongly between the two cases. We sought to quantify and compare these disparate neural and perceptual effects.

For the neural data, we measured the relative distance between pairs of image responses by measuring the Euclidean distance between normalized population response vectors (see *Methods: Perceptual and neural distance metrics*). Neural distances were relatively large between photographic images and scrambled conditions and relatively small between any two scrambled conditions. This is visually evident from the lighter cell values on the left column and bottom row of the plot in Figure 5A.

We next considered whether this categorical separation of photographic images from all scrambled conditions was also present in perceptual measurements of the same images. We first turned to two standard measures of image similarity that capture some simple aspects of perceptual similarity: 1) pixel-based Euclidean distance and 2) the structural similarity index (SSIM, Wang et al., 2004) (Figure 5B, left and center). As with neural distances, we measured images similarity distances between pairs of images, then averaged them over all pairs. Unlike V4 neural distances, both of these measures displayed a smooth gradation from fully scrambled images to partially scrambled images to photographic images. We concluded that the categorical separation of photographic images and scrambled images in the V4 population response could not be explained by these relatively simple properties of the images.

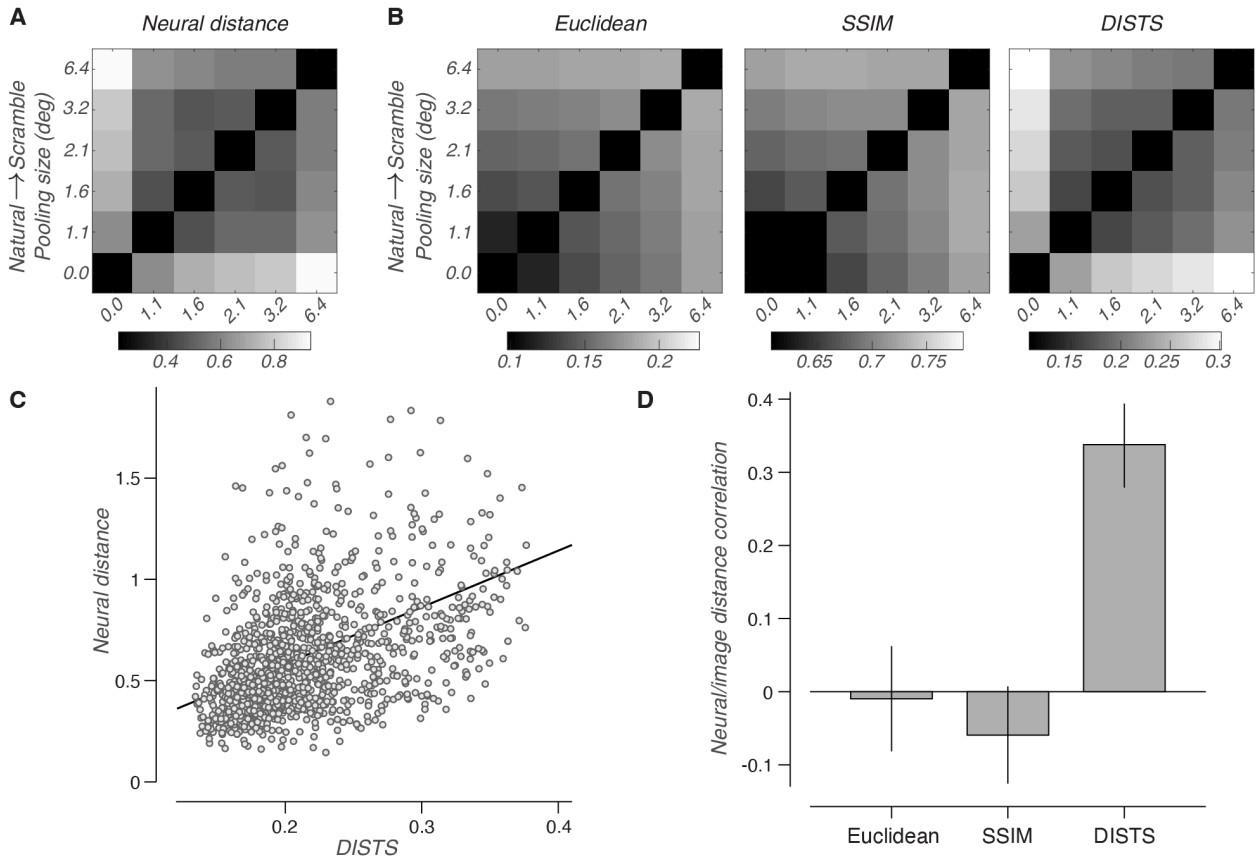

Figure 5. The image quality metric DISTS predicts the V4 population response. A) Neural distances, measured between pairs of V4 population responses. Distances were only measured between matched images (same family and shift) with different levels of scrambling. Pooling regions range from fully scrambled images (6.4 degrees) to photographic images (0.0 degrees). Colors denote Euclidean distances between pairs of normalized population vectors, averaged over all appropriate image pairs. The white banding on the left and bottom edges indicates that neural populations categorically separate photographic images from all scrambled conditions. B) Image similarity, measured between matched images with different levels of scrambling. We used three different distance metrics: 1) average RMS pixel distance (left), 2) SSIM (center), and 3) DISTS (right). Pixel distance and SSIM differ from neural distance, while DISTS appears similar. C) Neural distance vs. DISTS for each measured pair of images. For 20 image families and 4 shift conditions per family, we measured 30 image comparisons (2400 total pairs plotted). The linear best fit is plotted in black. D) Correlations between neural distances and the three image similarity metrics, averaged over resampled neural subpopulations (N=105 neurons). Error bars represent standard deviations across subpopulations. DISTS predicts neural distance, while pixel distance and SSIM do not.

Next, we turned to a more complex image similarity measurement: the Deep Image Image Structure and Texture Similarity metric (DISTS, Ding et al., 2022). Unlike pixel distance or SSIM, this metric was developed to be robust to image distortions that preserve texture identity, and was fit directly to human similarity judgements of distorted

and photographic images. Like our V4 measurements, the DISTS metric showed a sharp split between photographic images and all scrambling conditions (Figure 5B, right). Furthermore, individual image pair distances measured with DISTS were significantly correlated with neural distances measured in V4 (Figure 5C-D), while measurements using pixel distances or SSIM showed essentially no correlation (Figure 5D). These results show that V4 population responses are well predicted by DISTS, and thus reflect the sensitivity of human observers to small, unnatural distortions in photographic images.

Sensitivity to image scrambling emerges slowly

Finally, we examined how neural sensitivity to image scrambling evolved over the time course of the response. To this end, we retrained classifiers for both the photographic structure detection task and the image family discrimination task in a series of 50 ms bins that spanned the temporal extent of the response.

First, we looked at the time course of photographic structure detection. Specifically, we measured the cross-family discriminability (as in Figure 2E) between fully scrambled images and either photographic images or partially scrambled images (Figure 6A). Sensitivity was slightly lower in the smaller time bins compared with the full temporal extent of the response (d'=1.33 peak, vs. d'=1.48 for the full 350 ms response), but still exhibited a large difference between photographic image sensitivity (red curve) and sensitivity for intermediate levels of scrambling (maroon/blue curves). For conditions under which we could robustly extract a time course (scrambling regions of 1.60 degrees or smaller), we found a similar time to peak sensitivity for full photographic images (125 ms) and for moderately scrambled images (1.07 deg.=124 ms, 1.60 deg.=107 ms).

Next, we looked at the dynamics of pairwise image family discriminability (Figure 6B). Again, peak sensitivity was only slightly lower in the smaller time bins compared with the full temporal extent of the response (d'=3.24, vs. d'=3.80 for the full 350 ms response). At early times (~150 ms, near peak performance), discriminabilities showed a graded improvement between fully scrambled images, intermediate levels of scrambling, and photographic images. At later time points (~300 ms), discriminabilities for full and intermediate levels of scrambling moved closer together, while photographic image discriminability exhibited a larger separation from intermediate scrambling conditions. Despite peaking at different sensitivities, all levels of scrambling took similar amounts of time to reach half of their maximum performance (all between 71 ms and 80 ms).

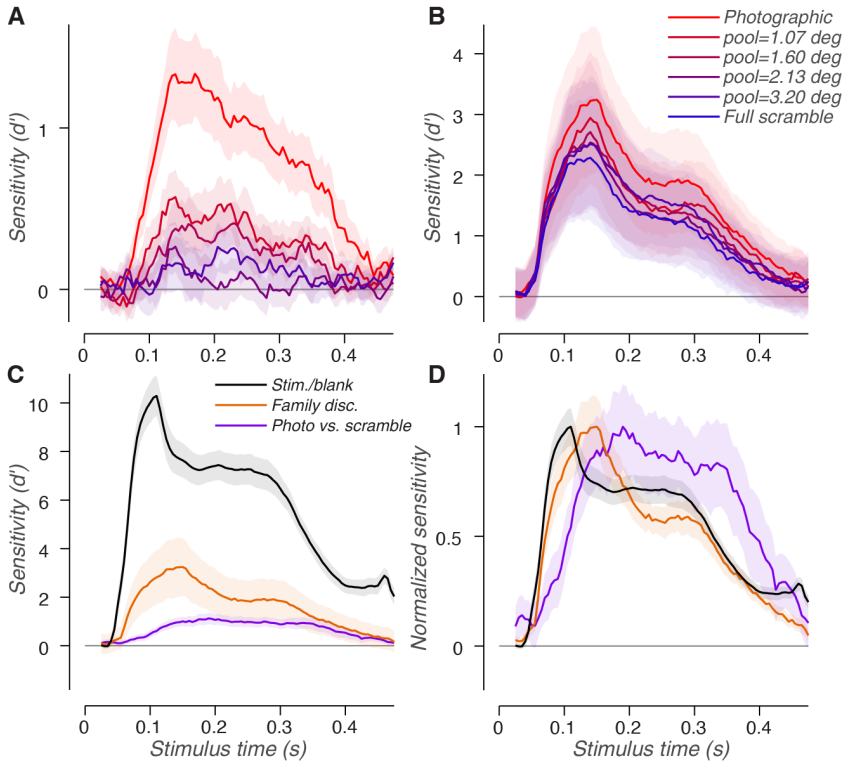

Figure 6: Sensitivities to different tasks emerge at different times. A) Discriminability between fully scrambled images and either photographic images (red) or intermediately scrambled images (other colors), plotted over time. Sensitivities were computed within 50 ms windows. Error bars represent standard deviations across subpopulations. B) Pairwise image family discriminability plotted over time. As in A), responses were calculated in 50 ms bins. Plots of different colors represent discriminations of photographic images (red), fully scrambled images (blue), or intermediately scrambled images (other colors). Error bars represent standard deviations across subpopulations. C) Comparison of sensitivity time courses for 3 different tasks: discrimination between image trials and blank trials (black), pairwise photographic family discrimination (orange), and photographic structure detection (purple). D) The 3 task traces from C) replotted as normalized sensitivities, such that peak sensitivity is 1. Different task sensitivities emerge at different times.

We next directly compared the time courses of these two tasks. In addition to photographic structure detection (Figure 6A) and image family discrimination (Figure 6B), we also measured the time course of discriminability between stimulus trials and blank trials. This can be thought of as a baseline measurement of V4's response to large changes in contrast.

Discriminabilities for these three tasks strongly differed in magnitude (Figure 6C). Stimulus-blank discriminability was more than twice as large as image family discriminability, which was again much larger than photograph vs. scramble discriminability. To compare the relative time courses of these tasks, we computed a *normalized sensitivity*: each task's sensitivity time course divided by its peak sensitivity (Figure 6D). The normalized curves show that stimulus-blank discriminability emerges first (half-max latency=65 ms), followed closely by the photographic family discrimination task (75 ms), and last (with substantial delay) by photographic vs. scrambled sensitivity (116 ms).

We estimated that stimulus/blank discriminability began rising in an analysis bin centered at 44 ms. Because the classification analysis is performed in 50 ms wide bins, this implies that stimulus information first emerges at the edge of this bin, 25 ms later, at 69 ms. This latency for V4 is consistent with previous reports (Lee et al., 2007; Oleskiw et al., 2018; Zamarashkina et al., 2020; Kim et al., 2022). We defined the “offset latency” as a time bin centered 200 ms later (at 245 ms). Activity after this point represents V4 responses after the image has left the screen. In all cases, we observed both robust firing rates and high classification sensitivities for at least 150 ms after this offset point. Notably, of the three tasks, the photographic structure detection tasks persisted at near-peak values for the longest, well past the point of stimulus offset. That is, family-general sensitivity was characterized by both a slow onset and slow decay, which suggests that it may arise from recurrent or feedback mechanisms.

Discussion

Natural images contain both objects (“things”) and materials (“stuff”) (Adelson and Bergen, 1991; Adelson, 2001). We created families of images that were matched in their global and local texture, and used them to show that population responses in V4 can reliably distinguish the presence or absence of object structure. This neural selectivity matches measurements from human perception (as simulated by an image similarity metric), emerges slowly, and persists through the full extent of the response. Photographic structure also improved the object recognition capabilities of the V4 population response (as measured by an image discrimination task) by increasing the dynamic range of firing rate responses.

Objects and textures

The images that we generated ranged between photographs, typically with centered objects, and matched textures. However, the distinction between objects and textures is an idealized one. Natural scenes typically contain mixtures of objects and textured surfaces at different locations and scales. Moreover, visual textures typically consist of collections of similar objects (e.g., a field of grass consists of many individual blades), and objects often contain textures surfaces or parts (e.g., dogs are covered with fur).

For this study, we selected photographic images that contained objects centered within natural backgrounds. In a sense, object boundaries within these images define a set of regions, either inside or outside the object, each with a relatively homogenous texture. One might ask why these images differ perceptually from the partially scrambled images, which are explicitly generated as a series of distinct subregions, each with a homogenous texture. One possible explanation may relate to the sharpness of the boundaries in each set of images. Photographic scenes typically contain sharp texture boundaries, while partially scrambled images are generated with much more gradual

texture boundaries. The fact that scrambled regions “leaked” into each other (and that photographic boundaries did not) may be the feature driving the scrambling modulation we observed in V4.

Object detection

V4 neuronal selectivities spanned a continuum between those preferring shape and those preferring texture, consistent with previous reports (Kim et al., 2019; Willeke et al., 2023). We found a relatively balanced distribution of these two populations, so that photographic images and scrambled images drove similarly strong responses when averaged across the V4 population. While response magnitude did not differentiate photographic and scramble images, response variance did – V4 neurons responded with a wider dynamic range to photographic images than to scrambled images (cf Rust and DiCarlo, 2010).

Some previous studies comparing object and texture responses have shown results that seem to differ from ours. First, two human fMRI studies have reported a stronger V4 response to photographic images than scrambled images (Movshon and Simoncelli, 2014; Long et al., 2018). This may be due to differences between single unit neurophysiology and fMRI. While the hemodynamic responses captured by BOLD fMRI are related to more direct measurements of neural activity, this relationship is not linear (Logothetis et al., 2001). In our data, V4 neural responses had a greater dynamic range for photographic images than for scrambled images, without a change in the mean. A nonlinear relationship between neuronal firing and BOLD could convert this difference in range into a difference in the mean fMRI signal.

Second, two studies using an image synthesis method similar to ours (TexForms) have reported that original images drive stronger neural responses than scrambled images (Long et al., 2018; Kramer et al., 2023). One possible reason for this difference with our study is that we used photographs with natural backgrounds, while the TexForm objects were cropped and placed on grey backgrounds. As a result, our photographic and scrambled image families had similar distributions of contrast within their centers and peripheries. In comparison, the TexForm object images likely had more contrast concentrated in their center, which may have made them more effective at driving V4 activity.

Finally, a study using a texture synthesis model based on a deep neural network (DNN) found that fMRI responses across the ventral stream, including areas V4 and IT, could not reliably distinguish between their photographic images and scrambled counterparts (Jagadeesh and Gardner, 2022). While these results are apparently consistent with ours, their DNN-based texture model generates images using statistics from the deepest layers of a DNN. These statistics predict firing rate responses in later visual

areas like IT (Yamins et al., 2014). Conversely, the Portilla-Simoncelli texture synthesis procedure only computes simple correlations of V1-like statistics, which are more closely associated with earlier areas like V2 (Portilla and Simoncelli, 2000; Freeman et al., 2013). The DNN-based statistics capture more characteristically “object-like” features for synthesis, and therefore create images sets that are less strongly differentiated than the image sets used in this study.

V4 responses can distinguish objects from textures. This distinction likely contributes to downstream processing in IT. IT responses are more strongly modulated by images of objects than by textures (Rust and DiCarlo, 2010), likely because texture image responses are more strongly suppressed in IT than in V4 (Lee et al., 2024). These results are consistent with psychophysical studies of perceptual crowding, which have shown that crowded visual scenes, like textures, are encoded with a compressed perceptual representation, particularly when compared to the representation of objects (Levi, 2008; Balas et al., 2009; Ziemba and Simoncelli, 2021). V4 is a key point in this perceptual process: human observers with more cortical area devoted to V4 are less likely to perceive objects as crowded within their surrounding texture (Kurzawski et al., 2024). Ventral stream processing suppresses responses to textures and elaborates representations of objects. V4 may be where the distinction between these two categories first emerges in an object-invariant manner.

Object and texture discrimination

Population responses in V4 give the first hint that the visual pathway is designed to recognize objects, a major computational goal of the ventral stream of visual cortex (DiCarlo et al., 2012). Not only could these responses discriminate between photographic images, but they also discriminated matched texture images that lack object-specific features. Roughly 1/3 of the object recognition signal in V4 relied on object-specific features in the images, while 2/3 relied on properties that were shared with the textures (Figure 3B). That is, simple image features such as V1-like spectral properties (like spatial frequency and orientation) and more complex, V2/V4-like “naturalistic” features (Freeman et al., 2013; Okazawa et al., 2015, 2016; Lee et al., 2024) carry the majority of the object recognition signal in V4. This ratio is likely a reflection of the fact that V4 is an early stage of object selectivity. Comparable measurements in downstream area IT have demonstrated an object recognition signal that is more strongly dependent on object-specific features in the images (Rust and DiCarlo, 2010).

Responses to partially scrambled images

For both classification tasks measured in this study, we observed that neural responses to partially scrambled images appeared to be “closer” to fully scrambled responses than

to photographic responses. This is consistent with previous observations that V4 responses are robustly driven by shape- and object-based representations (Pasupathy et al., 2020). Neuronal responses in V4 are selective for complex shape (Kobatake and Tanaka, 1994; Pasupathy and Connor, 1999, 2001) and lesions to area V4 profoundly disrupt form-processing behaviors (Merigan, 1996). Many neurons in V4 are selective for the sharpness of object edges (Oleskiw et al., 2018) and monocular cues for three-dimensional shape (Srinath et al., 2021), both of which are disrupted by image scrambling.

These neural results are mirrored by perceptual results, where human observers are highly sensitive to subtle disruptions in object closure (Elder et al., 2018) or small amounts of scrambling in object images, but not texture images (Wallis et al., 2019). Measuring and modeling human sensitivity to image distortion is an active problem in computer vision research (Lin et al., 2019). Inspired by this line of work, we found that DISTS, a computer vision model of image distortion perception (Ding et al., 2022), also predicts that small amounts of image scrambling should be easy to detect. Accordingly, we found that DISTS could partially account for V4's sensitivity to scrambling.

Response dynamics

The modulation of V4 responses by photographic image structure first emerges ~50 ms after stimulus onset, and grows slowly thereafter. This is consistent with many other studies that have described slowly-developing signals within V4. Some stimulus-driven properties, such as selectivity for complex contours (Yau et al., 2013) or shapes with blurred edges (Oleskiw et al., 2018), emerge more quickly than scrambled image modulation. Others, like V4's selectivity to a (different) set of perceptually-salient textural features (Kim et al., 2022) emerges over a time course that is similar to scrambled image modulation. All of these stimulus-driven effects typically emerge faster than those related to more "cognitive" processes, such as the onset of attentional modulation (Motter, 1994) and the "filling in" of occluded stimuli that is hypothesized to originate in prefrontal areas (Fyall et al., 2017).

V4's selectivity for object identity emerged relatively quickly (10 ms after stimulus onset), and depended strongly on the low- and mid-level features captured by the Portilla-Simoncelli model. This suggests that the object selectivity signal is primarily built up from feedforward signals originating in areas V1 and V2. Conversely, the relatively delayed time course of modulation due to image scrambling (between images matched in their low- and mid-level features) suggests that selectivity for high-level, object-selective image features may rely on recurrent processing within V4. Alternatively, the signal may rely on feedback from areas further down the ventral stream, such as posterior inferotemporal cortex (PIT) (Felleman and Van Essen, 1991), where object images modulate responses more strongly than matched texture images

(Rust and DiCarlo, 2010). Such feedback signals may propagate further upstream than V4 – it has recently been reported that some neurons in V1 also show a preference for photographic images relative to scrambled images (Chen et al., 2022; Kramer et al., 2023).

Conclusions

A compact way of describing the functional role of a visual area is to describe the visual attributes for which it is selective, and those for which it is invariant. In this framework, selectivities and invariances should become more complex for areas further along the visual ventral stream. Early in this stream, complex cells in V1 build a representation of local orientation that is invariant to changes in spatial position (Movshon et al., 1978; Vintch et al., 2015). Neurons in V2 show selectivity for texture statistics, and invariance across image transformations that preserve those statistics (Ziemba et al., 2016). At the end of the ventral stream, neurons in IT build selective representations of visual objects that are invariant to identity-preserving transformations like translation, rotation, and background context (Rust and DiCarlo, 2010; Hong et al., 2016). Here, we show that populations of neurons in V4 exhibit selectivity for both the presence and identity of objects. Somewhat counterintuitively, V4's selectivity for the presence of objects is invariant to changes in the specific object identity, and its selectivity for object identity is invariant to the presence or absence of objects in the image. That is because these representations live in independent subspaces. Building this "untangled" representation of object detection and object identification may be the primary goal of ventral stream processing (DiCarlo and Cox, 2007).

References

Adelson EH (2001) On seeing stuff: the perception of materials by humans and machines. In: Human Vision and Electronic Imaging VI, pp 1–12. SPIE.

Adelson EH, Bergen JR (1991) The Plenoptic Function and the Elements of Early Vision. In: Computational Models of Visual Processing, pp 3–20. MIT Press.

Balas B, Nakano L, Rosenholtz R (2009) A summary-statistic representation in peripheral vision explains visual crowding. Journal of Vision 9:13.

Buccino AP, Hurwitz CL, Garcia S, Magland J, Siegle JH, Hurwitz R, Hennig MH (2020) SpikeInterface, a unified framework for spike sorting. eLife 9:e61834.

Bushnell BN, Harding PJ, Kosai Y, Pasupathy A (2011) Partial Occlusion Modulates Contour-Based Shape Encoding in Primate Area V4. J Neurosci 31:4012–4024.

Chen X, Zhu S, Bai K, Xia R, Kong NCL, Norcia AM, Moore T (2022) Rapid Selectivity to Natural Images Across Layers of Primate V1. :2022.01.23.477422.

Chung JE, Magland JF, Barnett AH, Tolosa VM, Tooker AC, Lee KY, Shah KG, Felix SH, Frank LM, Greengard LF (2017) A Fully Automated Approach to Spike Sorting. Neuron 95:1381-1394.e6.

DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. Trends in Cognitive Sciences 11:333–341.

DiCarlo JJ, Zoccolan D, Rust NC (2012) How Does the Brain Solve Visual Object Recognition? Neuron 73:415–434.

Dickey AS, Suminski A, Amit Y, Hatsopoulos NG (2009) Single-Unit Stability Using Chronically Implanted Multielectrode Arrays. Journal of Neurophysiology 102:1331–1339.
Ding K, Ma K, Wang S, Simoncelli EP (2022) Image Quality Assessment: Unifying Structure and Texture Similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence 44:2567–2581.
Elder JH, Oleskiw TD, Fruend I (2018) The role of global cues in the perceptual grouping of natural shapes. Journal of Vision 18:14.
El-Shamayleh Y, Pasupathy A (2016) Contour Curvature As an Invariant Code for Objects in Visual Area V4. J Neurosci 36:5532–5543.
Felleman DJ, Van Essen DC (1991) Distributed Hierarchical Processing in the Primate Cerebral Cortex. Cerebral Cortex 1:1–47.
Freeman J, Simoncelli EP (2011) Metamers of the ventral stream. Nat Neurosci 14:1195–1201.
Freeman J, Ziemba CM, Heeger DJ, Simoncelli EP, Movshon JA (2013) A functional and perceptual signature of the second visual area in primates. Nat Neurosci 16:974–981.
Fukushima K (1980) Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybernetics 36:193–202.
Fyall AM, El-Shamayleh Y, Choi H, Shea-Brown E, Pasupathy A (2017) Dynamic representation of partially occluded objects in primate prefrontal and visual cortex. eLife 6:e25784.
Goodale MA, Milner AD (1992) Separate visual pathways for perception and action. Trends in Neurosciences 15:20–25.
Green DM, Swets JA (1966) Signal detection theory and psychophysics. Oxford, England: John Wiley.
Hong H, Yamins DLK, Majaj NJ, DiCarlo JJ (2016) Explicit information for category-orthogonal object properties increases along the ventral stream. Nat Neurosci 19:613–622.
Jagadeesh AV, Gardner JL (2022) Texture-like representation of objects in human visual cortex. Proc Natl Acad Sci U S A 119:e2115302119.
Josephs EL, Zhao H, Konkle T (2021) The world within reach: An image database of reach-relevant environments. Journal of Vision 21:14.
Kaas JH, Qi H-X, Stepniewska I (2022) Escaping the nocturnal bottleneck, and the evolution of the dorsal and ventral streams of visual processing in primates. Philosophical Transactions of the Royal Society B 377.
Kim T, Bair W, Pasupathy A (2019) Neural Coding for Shape and Texture in Macaque Area V4. J Neurosci 39:4760–4774.
Kim T, Bair W, Pasupathy A (2022) Perceptual Texture Dimensions Modulate Neuronal Response Dynamics in Visual Cortical Area V4. J Neurosci 42:631–642.
Kobatake E, Tanaka K (1994) Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. Journal of Neurophysiology 71:856–867.
Kramer LE, Chen Y-C, Long B, Konkle T, Cohen MR (2023) Contributions of early and mid-level visual cortex to high-level object categorization. :2023.05.31.541514.
Kurzawski JW, Qiu BS, Majaj NJ, Benson N, Pelli DG, Winawer J (2024) Human V4 size predicts crowding distance. :2024.04.03.587977.
Lee GM, Deliz CLR, Bushnell BN, Majaj NJ, Movshon JA, Kiorpes L (2024) Developmentally stable representations of naturalistic image structure in macaque visual cortex. Cell Reports 43.
Lee GM, Majaj NJ, Deliz CLR, Kiorpes L, Movshon JA (2025) Emergence of a contrast-invariant representation of naturalistic texture in macaque visual cortex. :2025.01.03.631258.
Lee J, Williford T, Maunsell JHR (2007) Spatial Attention and the Latency of Neuronal Responses in Macaque Area V4. J Neurosci 27:9632–9637.
Levi DM (2008) Crowding—An essential bottleneck for object recognition: A mini-review. Vision Research 48:635–654.

Lin H, Hosu V, Saupe D (2019) KADID-10k: A Large-scale Artificially Distorted IQA Database. In: 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX), pp 1–3.
Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A (2001) Neurophysiological investigation of the basis of the fMRI signal. Nature 412:150–157.
Logothetis NK, Sheinberg DL (1996) Visual Object Recognition. Annual Review of Neuroscience 19:577–621.
Long B, Yu C-P, Konkle T (2018) Mid-level visual features underlie the high-level categorical organization of the ventral stream. Proceedings of the National Academy of Sciences 115:E9015–E9024.
Majaj NJ, Hong H, Solomon EA, DiCarlo JJ (2015) Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. J Neurosci 35:13402–13418.
Merigan WH (1996) Basic visual capacities and shape discrimination after lesions of extrastriate area V4 in macaques. Visual Neuroscience 13:51–60.
Mishkin M, Ungerleider LG, Macko KA (1983) Object vision and spatial vision: two cortical pathways. Trends in Neurosciences 6:414–417.
Motter BC (1994) Neural correlates of attentive selection for color or luminance in extrastriate area V4. J Neurosci 14:2178–2189.
Movshon JA, Simoncelli EP (2014) Representation of Naturalistic Image Structure in the Primate Visual Cortex. Cold Spring Harb Symp Quant Biol 79:115–122.
Movshon JA, Thompson ID, Tolhurst DJ (1978) Receptive field organization of complex cells in the cat's striate cortex. J Physiol 283:79–99.
Nandy AS, Sharpee TO, Reynolds JH, Mitchell JF (2013) The Fine Structure of Shape Tuning in Area V4. Neuron 78:1102–1115.
Okazawa G, Tajima S, Komatsu H (2015) Image statistics underlying natural texture selectivity of neurons in macaque V4. Proc Natl Acad Sci USA 112:E351–E360.
Okazawa G, Tajima S, Komatsu H (2016) Gradual Development of Visual Texture-Selective Properties Between Macaque Areas V2 and V4. Cereb Cortex 27:4867–4880.
Oleskiw TD, Nowack A, Pasupathy A (2018) Joint coding of shape and blur in area V4. Nat Commun 9:466.
Pasupathy A, Connor CE (1999) Responses to Contour Features in Macaque Area V4. Journal of Neurophysiology 82:2490–2502.
Pasupathy A, Connor CE (2001) Shape Representation in Area V4: Position-Specific Tuning for Boundary Conformation. Journal of Neurophysiology 86:2505–2519.
Pasupathy A, Connor CE (2002) Population coding of shape in area V4. Nat Neurosci 5:1332–1338.
Pasupathy A, Popovkina DV, Kim T (2020) Visual Functions of Primate Area V4. Annual Review of Vision Science 6:363–385.
Portilla J, Simoncelli EP (2000) A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. International Journal of Computer Vision 40:49–70.
Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2:1019–1025.
Rust NC, DiCarlo JJ (2010) Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT. Journal of Neuroscience 30:12978–12995.
Sharpee TO, Kouh M, Reynolds JH (2013) Trade-off between curvature tuning and position invariance in visual area V4. Proceedings of the National Academy of Sciences 110:11618–11623.

Srinath R, Emonds A, Wang Q, Lempel AA, Dunn-Weiss E, Connor CE, Nielsen KJ (2021) Early Emergence of Solid Shape Coding in Natural and Deep Network Vision. Curr Biol 31:51-65.e5.
Steinmetz NA et al. (2021) Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. Science 372:eabf4588.
Tkačik G, Garrigan P, Ratliff C, Milčinski G, Klein JM, Seyfarth LH, Sterling P, Brainard DH, Balasubramanian V (2011) Natural Images from the Birthplace of the Human Eye. PLOS ONE 6:e20409.
Ungerleider LG, Mishkin M (1982) Two cortical visual systems. In: Analysis of visual behavior, pp 549–586. MIT Press.
Vintch B, Movshon JA, Simoncelli EP (2015) A Convolutional Subunit Model for Neuronal Responses in Macaque V1. J Neurosci 35:14829–14841.
Wallis TS, Funke CM, Ecker AS, Gatys LA, Wichmann FA, Bethge M (2019) Image content is more important than Bouma's Law for scene metamers Behrens TE, Herzog M, Herzog M, Caas J, eds. eLife 8:e42512.
Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13:600–612.
Willeke KF, Restivo K, Franke K, Nix AF, Cadena SA, Shinn T, Nealley C, Rodriguez G, Patel S, Ecker AS, Sinz FH, Tolias AS (2023) Deep learning-driven characterization of single cell tuning in primate visual area V4 unveils topological organization. :2023.05.12.540591.
Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci USA 111:8619–8624.
Yau JM, Pasupathy A, Brincat SL, Connor CE (2013) Curvature Processing Dynamics in Macaque Area V4. Cerebral Cortex 23:198–209.
Zamarashkina P, Popovkina DV, Pasupathy A (2020) Timing of response onset and offset in macaque V4: stimulus and task dependence. J Neurophysiol 123:2311–2325.
Ziemba CM, Freeman J, Movshon JA, Simoncelli EP (2016) Selectivity and tolerance for visual texture in macaque V2. Proc Natl Acad Sci USA 113.
Ziemba CM, Freeman J, Simoncelli EP, Movshon JA (2018) Contextual modulation of sensitivity to naturalistic image structure in macaque V2. Journal of Neurophysiology 120:409–420.
Ziemba CM, Perez RK, Pai J, Kelly JG, Hallum LE, Shooner C, Movshon JA (2019) Laminar Differences in Responses to Naturalistic Texture in Macaque V1 and V2. J Neurosci 39:9748–9756.
Ziemba CM, Simoncelli EP (2021) Opposing effects of selectivity and invariance in peripheral vision. Nat Commun 12:4597.