Responses of neurons in macaque V4 to object and texture images

Justin D. Lieber[1], Timothy D. Oleskiw[1,2], 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

Acknowledgements. We are grateful to Kahlia Gronthos, Sullivan Bacerdo, and Kaitlyn Holman for their assistance. Laura Palmieri and Manu Raghavan participated in some of the experiments and helped 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. J.D.L was supported in part by a Leon Levy Fellowship in Neuroscience.

Abstract

Humans and monkeys can effortlessly recognize objects in everyday scenes. This ability relies on neural computations in the ventral stream of visual cortex. The intermediate computations that lead to object selectivity are not well understood, but previous studies implicate V4 as an early site of selectivity for object shape. To explore the mechanisms of this selectivity, we generated a continuum of images between "scrambled" textures and photographic images of both natural and manmade environments, using techniques that preserve the local statistics of the original image while discarding information about scene and shape. We measured the responses of single units in awake macaque V4 to these images. On average, V4 neurons were slightly more active in response to photographic images than to their scrambled counterparts. However, responses in V4 varied widely both across different cells and different sets of images. An important determinant of this variation was the effectiveness of image families at driving strong neural responses. Across the full V4 population, a cell's average evoked firing rate for a family reliably predicted that family's preference for photographic over scrambled images. Accordingly, the cells that respond most strongly to each image family showed a much stronger difference between photographic and scrambled images and a graded level of modulation for images scrambled at intermediate levels. This preference for photographic images was not evident until ~50 ms after the onset of neuronal activity and did not peak in strength until 140 ms after activity onset. Finally, V4 neural responses seemed to categorically separate photographic images from all of their scrambled counterparts, despite the fact that the least scrambled images in our set appear similar to the originals. When these same images were analyzed with DISTS (Deep Image Structure and Texture Similarity), an image-computable similarity metric that predicts human judgements of image degradation, this same pattern emerged. This suggests that V4 responses are highly sensitive to small deviations from photographic image structure.

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 consists of a series of hierarchically connected visual areas: beginning in area V1, continuing through areas V2, then V4, before culminating in inferotemporal cortex (IT). Population responses of neurons in IT successfully discriminate objects (Pasupathy and Connor, 2002; Rust and DiCarlo, 2010) and predict human performance on object recognition tasks (Majaj et al., 2015). While the visual responses of neurons in IT have been well characterized, the intermediate computations that build these object selective responses are not yet well understood.

We recently described a selectivity for complex image statistics in V2 neurons that is not present in area V1 (Freeman et al., 2013; Ziemba et al., 2016, 2018). Using the Portilla-Simoncelli texture model, which is based on image statistics derived from a V1-like

representation (Portilla and Simoncelli, 2000), we generated images of "naturalistic" texture and spectrally matched noise. Neurons in V2 respond more strongly to images containing naturalistic structure than to matched noise images (Freeman et al., 2013), and this modulation emerges slowly over the course of 60-80 ms after response onset (Freeman et al., 2013; Okazawa et al., 2016; Ziemba et al., 2018).

The selectivities of neurons in area V4 are less well understood. One hypothesis is that object-centric representations are first constructed in V4 (Pasupathy et al., 2020). Many neurons in V4 respond robustly and selectively to object curvature (Pasupathy and Connor, 1999, 2001), with responses that are invariant to object position (El-Shamayleh and Pasupathy, 2016) and integrate global context about object occlusion (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 also responds more strongly to photographic images than to matched texture images synthesized using the Portilla-Simoncelli algorithm.

In this study, we asked how neurons in V4 respond to a continuum of images between photographic images and 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, respectively. Photographic images drove larger modulations in the V4 population response than scrambled images, and those modulations were delayed relative to the onset of neural activity. V4 responses were highly sensitive to even small amounts of image scrambling, such that partially scrambled conditions were categorically distinct from photographic images. This categorical separation between photographic and partially scrambled images is well captured by DISTS, an image quality assessment metric designed for invariance to changes in texture resampling.

Methods

Image generation

We chose a core set of 20 large photographic images taken from two databases. Half of these images were selected from photographs taken at a baboon habitat in Botswana (UPenn Natural Image Database, Tkačik et al., 2011), which approximate the evolutionary context of primate vision. The other half were selected from a photograph database of everyday objects in their natural context (Reachspace database, Josephs et al., 2021). These images are similar to some of the images that are typically used to train and evaluate deep neural network models of area V4 (Yamins et al., 2014). These source images were square cropped, when appropriate, to center objects within the frame. All source images were at least 800x800 pixels after cropping. Source images were then resized to a standard size of 1280x1280 pixels for image analysis and synthesis. From each source image, we then cropped four distinct 512x512 pixel regions, centered at locations +/-56 pixels horizontally and vertically from the center of

the source image. These images substantially overlap in their content and, when presented to the subject, represent a relative "shift" of the underlying pixels by a distance of 1.4 deg.

To generate scrambled images, we used an adaptation of the Portilla & Simoncelli texture synthesis model that measures and synthesizes texture images within localized 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. Then, within each pooling region, we measured a set of texture statistics (Portilla and Simoncelli, 2000). Specifically, we processed the image with a multi-scale, multi-orientation bank of filters (4 orientations, 4 scales), then computed both the linear responses and energy responses of each filter. We then computed the pairwise product of these responses at different positions, orientations, and scales. Finally, we computed weighted averages of these products within each pooling region, resulting in a set of correlation statistics within each region.

We synthesized new "scrambled" images by initializing an image with Gaussian white noise, then iteratively adjusting the pixels of that image to match the measured texture statistics within each subregion. The texture synthesis method uses periodic boundary conditions, measuring texture statistics toroidally across the edges of images. To prevent edge artifacts from appearing within the synthesized images, between iterations of the synthesis we "reset" the edge regions of the image (all pixels outside a circular vignette) to 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, so that only "scrambled" portions of the image remained.

For each "shifted" image, 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 images, giving a total of 6 distinct pooling region conditions. We refer to the set of 24 images derived from an original photograph (4 shifts x 6 pooling regions) 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 6-8 consecutive presentations, after which the animal received a juice reward.

Under general anesthesia, we implanted the animal 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 that were isolated and 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. Single units were isolated from multi-site recordings using KiloSort 2.5 (Kilosort 2.5, Steinmetz et al., 2021) followed by manual curation of well-isolated spikes (Phy, https://phy.readthedocs.io/).

Firing rate analyses

For all cells, we computed firing rates for all spikes within 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 by the mean firing rate response to all non-blank stimulus conditions.

Stimulus-dependent fractional variance: 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 stimulus-dependent fractional variance 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 the stimulus-dependent variance as the variance across each stimulus's firing rate averaged across repeated trials. Stimulus-dependent fractional variance 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. For each cell, we averaged firing rates over all image family and shift conditions to find a single rate for both the photographic or fully scrambled conditions. We compute modulation index as the difference between these rates, divided by their sum. For population-level modulation indices, we averaged rates over families, shifts, and neurons before computing the index.

Rate-modulation correlation: For each neuron, we quantified the strength of the relationship between an image family’s total response and its relative modulation of photographic vs. scrambled images as a rate-modulation correlation. 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 responses evoked by photographic images and fully scrambled images had different relationships to a neuron's preferences for image family. We were concerned that simply ranking each neuron's preferred image families might produce overfitting due to trial-to-trial fluctuations in firing rate. Overfit rankings could potentially overstate the difference in response magnitude between image families. 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). We computed average evoked firing rates over the photographic and fully scrambled conditions for each training set, then used these rates to rank image families. We then applied that ranking to the evoked firing rates of the test sets when plotting the data. 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.

Population latency: To compute the onset of neural activity or modulation, we fit a rectified linear function (using non-linear least squares optimization, lsqcurvefit() in Matlab) for a period of time from before onset to the peak of activity/modulation:

$$r(t) = r_b + m * [t - l]^+$$

where the three fit parameters are l, the onset latency, m, a slope parameter, and r_b, the baseline rate/modulation.

Perceptual and neural distance metrics

We wondered whether any well-established image measurements could predict the patterns of response we observed in V4. We turned to a set of image similarity metrics, which are most commonly used in practical applications to quantify the distance between an original image and a corrupted counterpart (e.g. to measure the quality of a lossy file format). To this end, we computed a set of pairwise distances between conditions for both the V4 population response, and for a set of image similarity metrics. In all cases, we only computed distances within the set of 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 original photographic image. Over 20 image families and 4 shifts per family, this resulted in 2,400 total distances.

Neural distance: For each image, we created a 134-neuron long 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), and Deep Image Structure and Texture Similarity (DISTS). 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 (Wang et al., 2004) using the *ssim()* function in Matlab, with all exponents set to 1. For the DISTS metric (Ding et al., 2022), we used a Matlab implementation of the algorithm provided by the authors (https://github.com/dingkeyan93/DISTS).

Results

We recorded the responses of 134 single units from area V4 of one macaque monkey. We chose 20 images from two natural image databases, 10 from the “Birthplace of the Eye” image set, containing evolutionarily relevant photographs taken at a primate reserve in Botswana (Tkačik et al., 2011), and 10 from the “Reachspace” image set (Josephs et al., 2021), containing images of objects in context in manmade scenes, with appropriate lighting and shadows. For each of these images, we created a larger “family” of 24 images, consisting of 4 discrete image shifts (Figure 1B) and scrambling at 5 distinct levels. Scrambling levels were chosen to transition smoothly between fully scrambled images and the original photographic images, in approximately equal perceptual increments (Figure 1C).

Scrambled image modulation is heterogeneous across a population of V4 neurons

V4 neurons responded well to both scrambled and photographic images (122/134 cells significantly responsive relative to baseline firing rates, $p<0.05$, permutation test). Most cells also responded differently to scrambled and photographic images. We captured this modulation by measuring the fraction of stimulus dependent variance (see Methods, Stimulus-dependent fractional variance). Stimulus dependent fractional variances were significantly greater than in permuted controls (median 0.23±0.13 MAD, 113/134 $p<0.05$, permutation test). The sign and magnitude of modulation due to image scrambling was heterogeneous across the population. Some cells preferred photographic images over scrambled images (Figure 2A-B), while others were indifferent (Figure 2C), or preferred scrambled images (Figure 2D). Averaging the responses across the full V4 population revealed weak modulation across image families and scrambling conditions (Figure 2E).

We defined a modulation index (MI) as the difference between firing rate responses to photographic and scrambled images, divided by their sum. On average, the population showed a weak preference for photographic images over scrambled images (MI mean = 0.01). Individual cells showed different amounts of modulation (Figure 3A). Overall, 38/134 cells were significantly positively modulated, and 26/134 cells were significantly negatively modulated ($p<0.05$, permutation test).

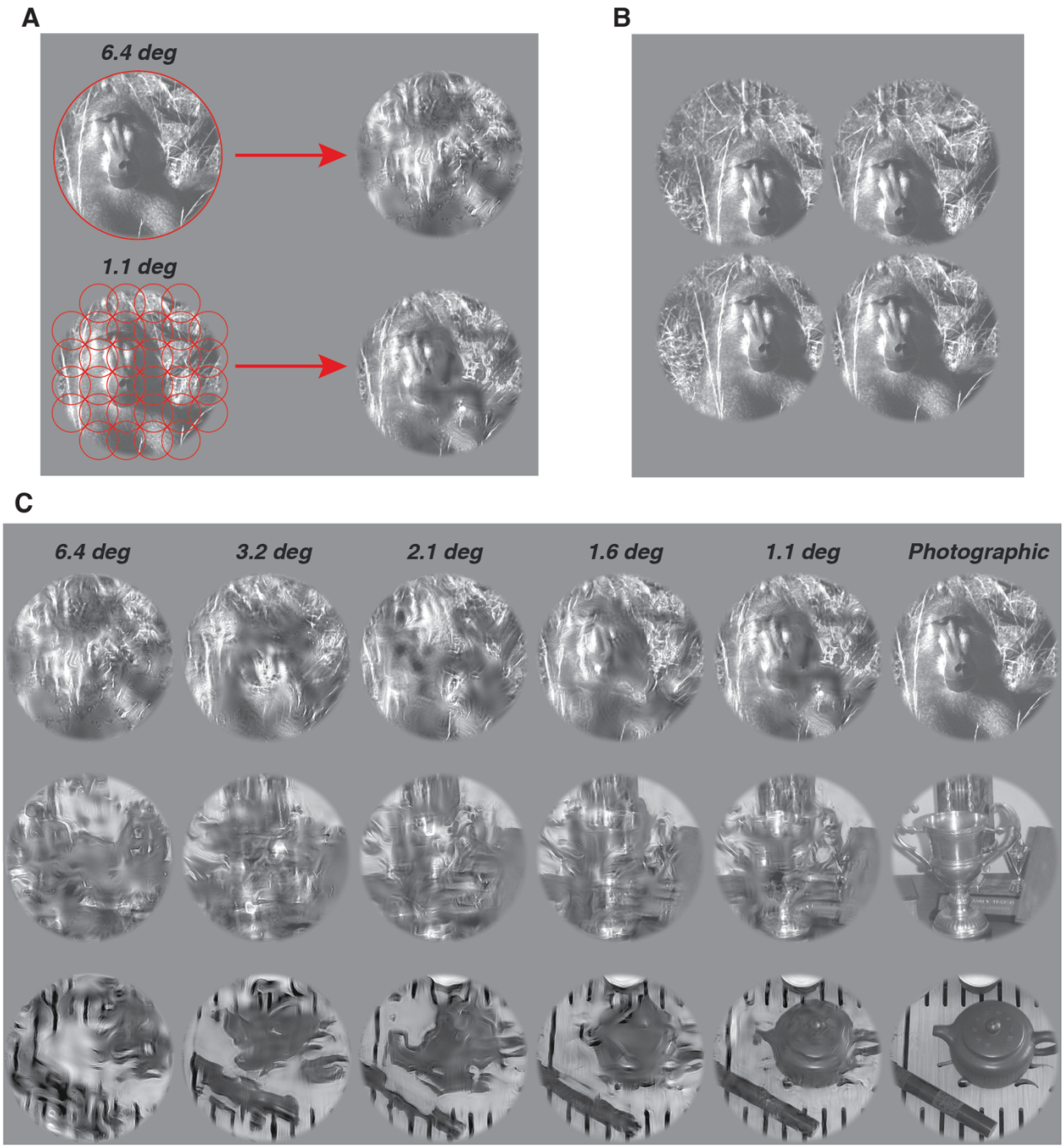

Figure 1. Creating a continuum of images from scrambled textures to photographic images. A) We used the Portilla-Simoncelli texture model to scramble image features within a local pooling region (red circle). When a single pooling region covered the image (top) the image was fully scrambled into a texture. When multiple smaller pooling regions covered the image (bottom) the resulting image maintains some global "object-like" properties, while local features are scrambled. B) For each source image, we cropped four distinct "shifts." These images contain the same central features, but are not matched pixel for pixel. 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 transition from scrambled textures to photographic images of objects.

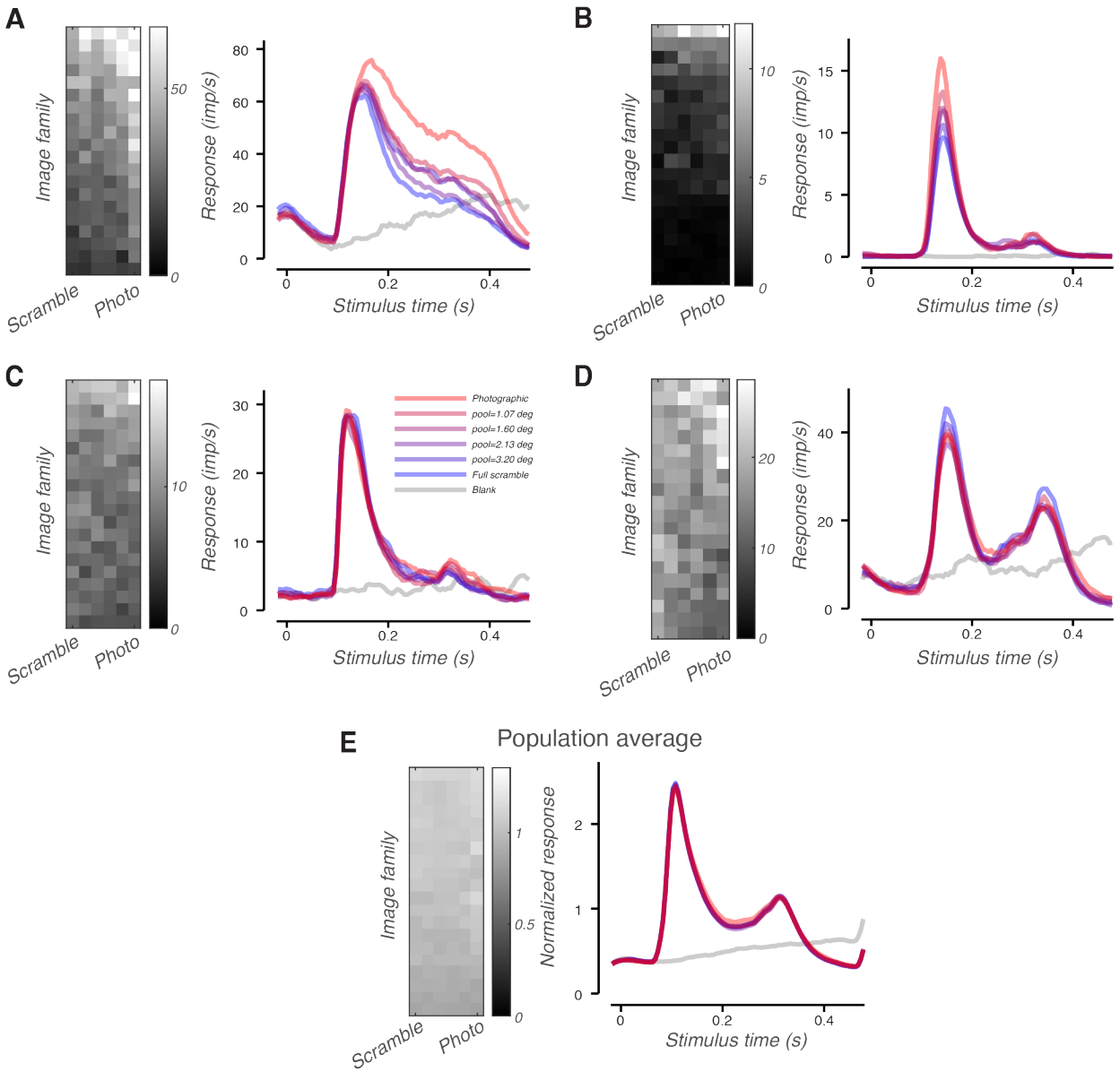

Figure 2: Responses of V4 neurons to photographic and scrambled images. A-D) Pixelated images on the left of each panel show the average firing rate response for all 20 image families and 6 scramble conditions. Pixel intensity indicates average firing rate over 4 image shifts. Image families are ordered from highest to lowest by average firing rate per family. Curves on the right of each panel show smoothed and averaged firing rates as peri-stimulus time histograms (PSTHs). While some neurons showed clear positive modulation of photographic images relative to scrambled images (A: MI=0.25, B: MI=0.18), others showed no obvious modulation (C: MI=0.01), or negative modulation (D: MI=-0.02). E) On the left, the population average normalized response for all image families and scramble conditions. Image families are sorted from highest to lowest normalized response. On the right, individual cell PSTHs were normalized and averaged over the population to produce a PSTH. The population response shows only very weak modulation on the average.

Image families based on natural scene photographs (“Birthplace of the Eye” image set) typically had weaker modulation than families based on photographs of manmade scenes (“Reachspace” image set). We measured modulation strength (both positive and negative) by taking the absolute value of the modulation index. Modulation strength for natural scenes (abs(MI)=0.040±0.025 MAD across cells) was typically weaker than that for manmade scenes (abs(MI)=0.061±0.035 MAD across cells), and significant for the population (population test: p<0.001, 42/134 individual cells significantly larger at p<0.05, permutation tests).

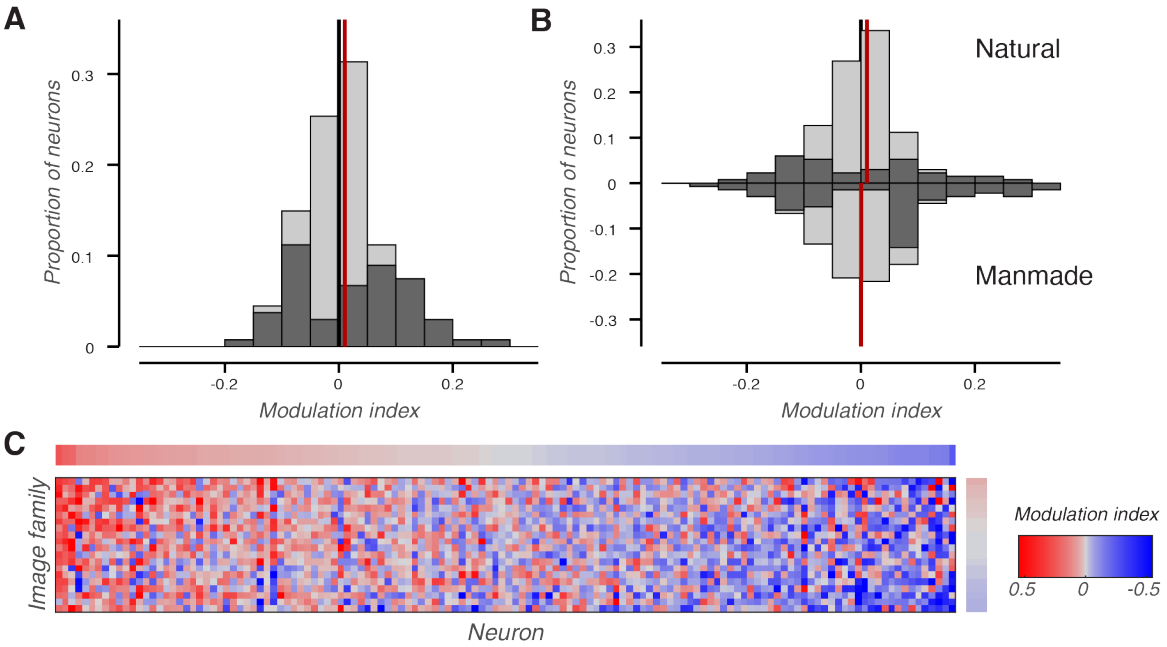

Figure 3: Modulation by image scrambling is heterogeneous across V4 neurons. A) Distribution of the modulation indices of individual neurons. Neurons with modulation indices that are significantly different from 0 are labeled in black (p<0.05). Nearly as many neurons were significantly positively modulated (N=36, p<0.025) as significantly negatively modulated (N=25, p<0.025). B) Distributions of modulation indices of individual neurons, split for images of natural scenes (top, Birthplace of the Eye image set) and images of manmade scenes (bottom, Reachspace image set). Neurons with modulation indices that are significantly different from 0 are labeled in black (p<0.05). C) The modulation index for all image families and cells. Both families and cells are ordered by their average modulation index (top to bottom, and left to right, respectively). The marginal color bars show the neuron-averaged difference for individual image families (right) or the family-averaged difference for individual neurons (top).

Scramble modulation varies with image family

Figure 3C reveals that for most individual neurons, the strength and sign of scramble modulation varied across different image families. Individual neurons sometimes exhibited significant modulation for a subset of image families, while not showing significant modulation in their average response to all families. 96/134 neurons had at least one family that was significantly positively modulated, 91/134 had at least one family that was significantly negatively modulated, and 68/134 had at least one of each

($p<0.05$, permutation test). The population also contained 15/134 neurons in which no family was significantly modulated ($p<0.05$, permutation test).

We next looked for a link between each neuron's preferred stimuli and the image families that most strongly modulated the cell. Neurons in V4 are tuned, at least in part, to the features that are matched between the scrambled and photographic images. Accordingly, we expected that the photographic images that most strongly drove neural responses in a given cell would be predictive of the scrambled images that drove strong responses. Consistent with previous reports (Long et al., 2018; Kramer et al., 2023), we found photographic image responses to be correlated with responses to scrambled members of the same image family (median correlation = 0.28 ± 0.14 MAD across all neurons). 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.

Based on the conjecture that stronger responses might be associated with stronger effects, we hypothesized that scrambled image modulation might be stronger for the image families that most strongly drove an individual cell. This turned out to be true, both in cells that were (on average) positively modulated (Figure 4A & 4C) and in those that were (on average) negatively modulated (Figure 4B & 4D). Across individual neurons, image family response correlated with image modulation (median correlation = 0.23±0.19 MAD, $p<0.001$, permutation test).

We summarized these effects at the population level by averaging each neuron's responses, rank-ordered by how strongly each family drove that cell (see Methods, Image family ranking). The families evoking the strongest responses were, on average, positively modulated. Additionally, the families driving the weakest responses were, on average, negatively modulated (Figure 4E). These results suggest that photographic images drive a wider dynamic range of responses than scrambled images. To directly measure this, for each cell we computed the coefficient of variation of firing rates as a measure of dynamic range, across either photographic images or fully scrambled images (Figure 4F). Photographic images typically drove larger amounts of variation than scrambled images (100/134 cells with larger variation for photographic images, 52/134 significantly larger, 15/134 significantly smaller at $p<0.05$, permutation test).

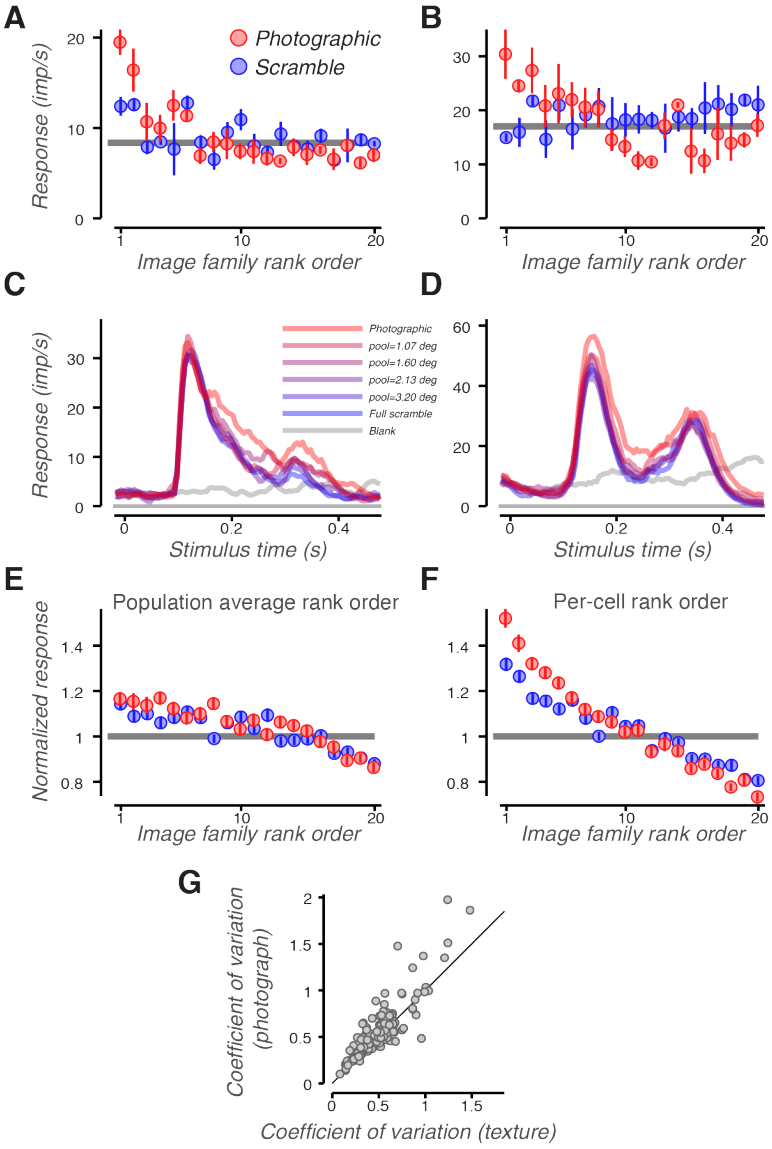

Figure 4. V4 responses to photographic images have a larger dynamic range than responses to scrambled images. A-B). For two individual neurons (neurons C & D from Figure 2), we ranked image families by how strongly they drove firing rate responses. We used a cross-validated procedure to rank image families (see Methods) to avoid overstating differences in image family responsiveness. Families that evoked strong responses also tended to show strong positive modulation. Conversely, families that evoked weaker responses were often negatively modulated. C-D) PSTHs for the same two neurons, computed only for the top quarter of families that most strongly drove firing rate responses. For this subset of images both cells were positively modulated (same cells as Figure 2 C-D, C: MI=0.17, D: MI=0.18). E) For the population averaged normalized response, we ranked image families by how strongly they drove responses. F) The population average over the rank order plots of individual V4 cells (computed as in A-B). Strongly driving families are positively modulated and weakly driving families are negatively modulated. G) The coefficient of variation of firing rates across all photographic images, plotted against the coefficient of variation across scrambled texture images.

V4 scrambled image modulation emerges slowly

Neuronal selectivities in area V4 emerged at different times after stimulus onset. To determine the timing of scrambled image modulation, we focused on a reduced subset of neurons (34 neurons most strongly, positively modulated) and image families (each neuron's 5 preferred families ranked by image family drive, as above). The resulting PSTH (Figure 5A) shows that even among this strongly modulated subset, scrambled image modulation emerged slowly. For this population, the stimulus-evoked firing rate first differed from baseline 73 ms after stimulus onset ("response onset"), but signals from different scrambling conditions only began to diverge 52 ms later (125 ms after stimulus onset). This modulation grew slowly, reaching a maximum 140 ms after response onset (213 ms after stimulus onset). Modulation remained significantly

positive during the offset response (at 273 ms) and for at least 200 ms after, partially overlapping with the onset of the following stimulus. The slow onset and persistent nature of this modulation signal suggest that it may reflect the action of recurrent circuits.

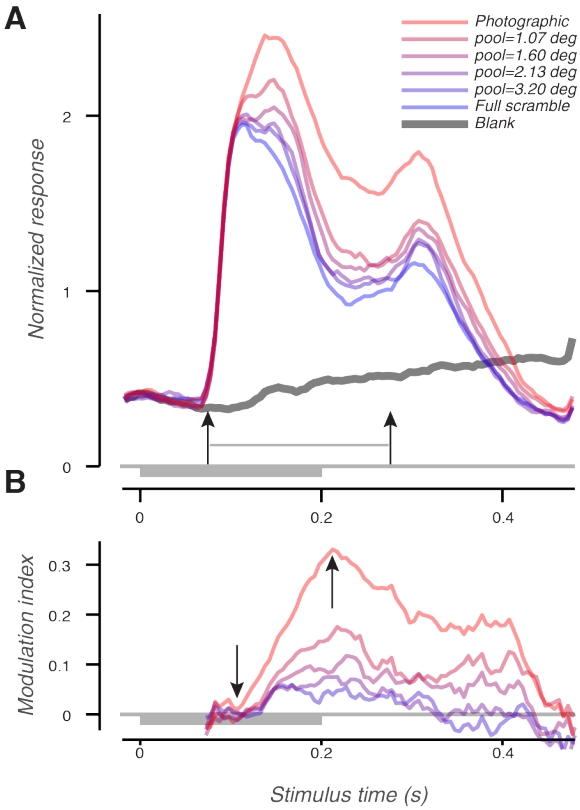

Figure 5. Modulation dynamics. A) Population averaged PSTHs for a subset of the neural population and a subset of the image set. Responses were averaged for the quarter of cells (34) that were most strongly, positively modulated. For each neuron, we took the top quarter of image families (5), ranked by how strongly they drove responses, averaged over both photographic and scrambled images. The onset of neural activity (at 73 ms) and beginning of the offset response (at 273 ms) are marked with arrows. Even among this strongly modulated subpopulation, modulation due to scrambling emerged 50 ms after the onset of neural activity. The uptick in activity beginning at 273 ms is due to a strong stimulus offset response from a subset of the neural population. B) Modulation indices between each scramble condition and the fully scrambled textures, at each point in time. Scramble modulation for photographic images first emerges at 125 ms (left arrow) and peaks at 213 ms (right arrow), and persists well after the response to stimulus offset.

An image similarity metric predicts V4 responses to photographic and scrambled images

Figure 5 also illustrates that, among this subset, the average modulation for photographic images (MI=0.17) is nearly twice as large as the modulation for any other intermediate scrambled condition (all MI<=0.091). To confirm that this effect was present among the larger population of cells and image families, we computed a population-level distance metric between the responses of pairs of images (Figure 6A). V4 population neural distances were consistently larger between photographic images and any scrambled condition than between different scrambled conditions – this is visually evident from the lighter cell values on the right and bottom margins of the plot in Figure 6A. This result was unexpected, as we had chosen intermediate scramble levels to approximate equally distinguishable increments (see, for example, Figure 1C). We confirmed that low-level properties of the image smoothly spanned the continuum between photographic and fully scrambled images using two image analysis metrics: pixel-based distance (Figure 6B) and the structural similarity index (SSIM, Wang et al., 2004) (Figure 6C). These metrics confirmed that low-level properties of the images transitioned smoothly among all the scrambling conditions and the photographic images.

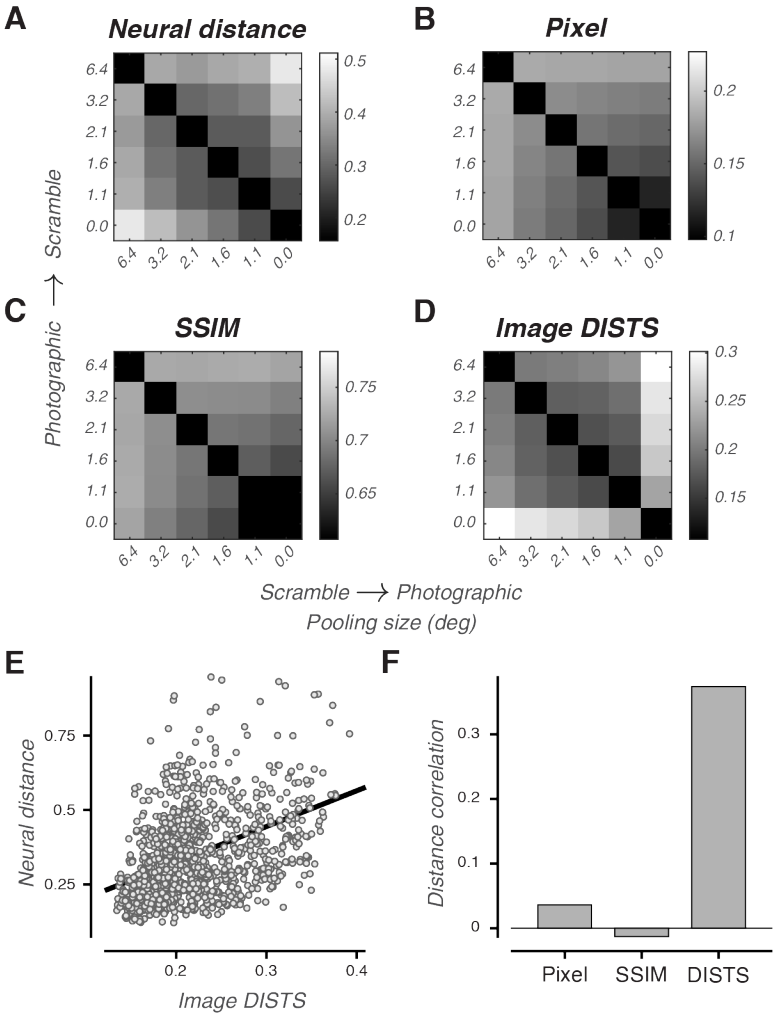

Figure 6. DISTS but not SSIM or pixel distance predicts the V4 population response to image scrambling. A-D) Pairwise, average distances for different pooling region sizes, for four different distance metrics. Large pooling region sizes correspond to larger amounts of scrambling (up to 6.4 degrees, a full scrambled image), while small sizes correspond to smaller amounts of scrambling (down to 0.0 degrees, the original photographic image). The first distance is a neural metric: the Euclidean distance between two normalized population vectors (A). The next three distance metrics derive from image analyses methods - average RMS pixel distance (B), the structural similarity index (SSIM) (C), or the Deep Image Structure and Texture Similarity Metric (D). DISTS and neural distance, but not pixel distance or SSIM, categorically separate photographic images from all scrambling conditions. E) Neural distance vs DISTS for individual pairs of images. For each of the 20 image families and 4 shift conditions (80 total photographic images), we measured distances between all 6 scramble conditions (30 comparisons per source image, 2400 pairs total). F) Correlation between individual condition neural distances and the three image similarity metrics. DISTS predicts neural distance, while pixel distance and SSIM do not.

One explanation for this discrepancy is that V4 neuronal responses are invariant to the precise arrangement of local features within images of texture, but not within images of objects. This response property would make responses to partially scrambled images more similar to each other, and more distinct from responses to photographic images. Human perception of texture images is thought to rely on a local computation of "summary statistics" (Balas et al., 2009; Greenwood et al., 2009; Freeman and Simoncelli, 2011; Rosenholtz et al., 2012; Ziemba and Simoncelli, 2021), such that different samples of the same texture appear nearly identical, even if the precise arrangement of their local features does not match.

To assess this possibility we used DISTS (Ding et al., 2022), a recently developed image similarity metric that was 1) designed to be robust to image variance that preserves texture identity, and 2) directly fit to human judgements of similarity between

distorted images and photographic images. Like our V4 measurements, the DISTS metric also showed a sharp split between photographic images and all scrambling conditions (Figure 6D). Furthermore, individual image pair distances measured with DISTS were significantly correlated with neural distances measured in V4 (Figure 6E), while measurements using pixel distances or SSIM showed essentially no correlation (Figure 6F). In short, DISTS could predict neural population responses to scrambled images.

Discussion

Our results show that neuronal responses in V4 are more strongly modulated by photographic images than by matched, scrambled images, even for very modest levels of scrambling. This modulation emerges slowly, and persists throughout the period over which neural activity is elevated.

Natural scene processing

Our results support the idea that some V4 responses are tuned to the characteristic properties of natural scenes. V4 responses are slightly stronger to photographic images than scrambled images, consistent with previous neurophysiology (Kramer et al., 2023) and human fMRI (Movshon and Simoncelli, 2014; Long et al., 2018). Populations of V4 neurons also more robustly classify photographic images than scrambled images (Rust and DiCarlo, 2010), consistent with our observation of a greater dynamic range of response to photographic images than scrambled images. At the individual neuron level, we find preferences for both photographic and scrambled images. This may be related to the previously-reported continuum of V4 responses from "shape-like" to "texture-like" (Kim et al., 2019; Willeke et al., 2023).

Jagadeesh and Gardner, (2022) report that human fMRI signals could not be used consistently to distinguish photographic images from paired scrambled images. However, their texture synthesis method used statistics from the late layers of a deep network model of object recognition, which are more likely to be computed in later visual areas like V4 and IT (Yamins et al., 2014). In contrast, the Portilla-Simoncelli textures are synthesized based on statistics likely to be captured in V2 (Portilla and Simoncelli, 2000; Freeman et al., 2013), and measurements of human fMRI using these textures have consistently shown modulation between photographic and scrambled images (Movshon and Simoncelli, 2014; Long et al., 2018).

The seemingly-categorical difference between photographic and even slightly scrambled images is consistent with previous observations that object-based representations drive robust neural responses in V4 (Pasupathy et al., 2020). Neurons in V4 are selective for complex shapes (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 to 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.

We hypothesize that V4's categorical separation of photographic image responses from mildly scrambled image responses may be a signature of an image quality computation. Although images scrambled with small pooling regions may appear perceptually similar to their photographic image counterparts, and may be relatively close in terms of pixel distance (Figure 5A), human observers are sensitive to scrambled image modulation even when pooling regions are very small (Wallis et al., 2019). We find that DISTS, an explicit model of this perceptual sensitivity, can partially account for how the V4 population responds to a wide array of image comparisons.

Dynamics

The modulation of V4 responses by photographic image structure emerges ~50 ms after stimulus onset, and grows slowly. This is also 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 the V4 population's selectivity to complex, perceptually-salient features of texture, emerge over a time course (Kim et al., 2022) similar to that of 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).

The delayed time course of scrambled image modulation may be a consequence of recurrent processing within V4, or feedback from areas further down the ventral stream, such as posterior inferotemporal cortex (PIT) (Felleman and Van Essen, 1991). 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

We conceived these experiments as a way to bridge representations early in the visual pathway – of what Adelson (2001) termed "stuff" – to later areas in which neurons show selective responses to images of natural objects – to Adelson, "things". Our results show that neurons in V4 do indeed respond selectively to images of objects, but they do so in an unconventional way. Rather than merely firing more spikes to object images, their responses more strongly differentiate images of preferred and non-preferred objects. This increased dynamic range allows V4 to provide more information about objects (Rust & DiCarlo, 2010), using this unconventional signaling strategy.

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. Available at: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/4299/0000/On-seeing-stuff--the-perception-of-materials-by-humans/10.1117/12.429489.full

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

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 Available at: https://www.biorxiv.org/content/10.1101/2022.01.23.477422v1 [Accessed October 24, 2023].

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

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.

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.

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 Rust N, ed. eLife 6:e25784.

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

Greenwood JA, Bex PJ, Dakin SC (2009) Positional averaging explains crowding with letter-like stimuli. Proceedings of the National Academy of Sciences 106:13130–13135.

Jagadeesh AV, Gardner JL (2022) Texture-like representation of objects in human visual cortex. Proceedings of the National Academy of Sciences 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 Available at:

https://royalsocietypublishing.org/doi/10.1098/rstb.2021.0293 [Accessed February 5, 2024].

Kaas JH, Qi H-X, Stepniewska I (2021) Escaping the nocturnal bottleneck, and the evolution of the dorsal and ventral streams of visual processing in primates. Phil. Trans. R. Soc. B377: 20210293.https://doi.org/10.1098/rstb.2021.0293

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 Available at: https://www.biorxiv.org/content/10.1101/2023.05.31.541514v1 [Accessed June 6, 2023].

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 feature selective memory and pop-out in extrastriate area V4. J Neurosci 14:2190–2199.

Movshon JA, Simoncelli EP (2014) Representation of Naturalistic Image Structure in the Primate Visual Cortex. Cold Spring Harb Symp Quant Biol 79:115–122.

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.

Rosenholtz R, Huang J, Raj A, Balas BJ, Ilie L (2012) A summary statistic representation in peripheral vision explains visual search. Journal of Vision 12:14.

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.

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. Current Biology 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.

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. 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 Available at: https://www.biorxiv.org/content/10.1101/2023.05.12.540591v1 [Accessed May 17, 2023].

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.

Ziemba CM, Freeman J, Movshon JA, Simoncelli EP (2016) Selectivity and tolerance for visual texture in macaque V2. Proc Natl Acad Sci USA 113 Available at: https://pnas.org/doi/full/10.1073/pnas.1510847113 [Accessed May 24, 2022].

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, Simoncelli EP (2021) Opposing effects of selectivity and invariance in peripheral vision. Nat Commun 12:4597.