Behavioral/Systems/Cognitive

Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT

Nicole C. Rust[1,2,3] **and James J. DiCarlo**[1,2]

[1]McGovern Institute for Brain Research and [2]Department Brain and Cognitive, Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, and [3]Department Psychology, University of Pennsylvania, Philadelphia, Pennsylvania 19104

Our ability to recognize objects despite large changes in position, size, and context is achieved through computations that are thought to increase both the shape selectivity and the tolerance ("invariance") of the visual representation at successive stages of the ventral pathway [visual cortical areas V1, V2, and V4 and inferior temporal cortex (IT)]. However, these ideas have proven difficult to test. Here, we consider how well population activity patterns at two stages of the ventral stream (V4 and IT) discriminate between, and generalize across, different images. We found that both V4 and IT encode natural images with similar fidelity, whereas the IT population is much more sensitive to controlled, statistical scrambling of those images. Scrambling sensitivity was proportional to receptive field (RF) size in both V4 and IT, suggesting that, on average, the number of visual feature conjunctions implemented by a V4 or IT neuron is directly related to its RF size. We also found that the IT population could better discriminate between objects across changes in position, scale, and context, thus directly demonstrating a V4-to-IT gain in tolerance. This tolerance gain could be accounted for by both a decrease in single-unit sensitivity to identity-preserving transformations (e.g., an increase in RF size) and an increase in the maintenance of rank-order object selectivity within the RF. These results demonstrate that, as visual information travels from V4 to IT, the population representation is reformatted to become more selective for feature conjunctions and more tolerant to identity preserving transformations, and they reveal the single-unit response properties that underlie that reformatting.

Introduction

Although our ability to identify individual objects invariant to position, size, and visual context may appear effortless, it is a tremendously complex computational challenge. The crux of the object recognition problem lies in the ability to produce a representation that can selectively identify individual objects in a manner that is essentially tolerant ("invariant") to changes in position, size, and context (Riesenhuber and Poggio, 1999; DiCarlo and Cox, 2007). From a computational perspective, constructing a representation that is either highly selective or highly tolerant is trivial; the challenge is to build a system that can produce a representation that is simultaneously selective and tolerant.

We do not fully understand how the brain accomplishes this task, but the solution is thought to be implemented through gradual increases in both selectivity and tolerance as signals propagate through the ventral visual stream [which includes the retina, lateral geniculate nucleus, visual cortical areas V1, V2, and V4, and inferior temporal cortex (IT)]. Evidence for gradual increases in selectivity is suggested by tuning for stimuli more complex than simple line segments in V2, V4, and posterior IT (Gallant et al., 1993; Pasupathy and Connor, 1999; Brincat and Connor, 2004; Anzai et al., 2007), as well as IT neurons that appear to be highly selective for complex objects (Desimone et al., 1984; Logothetis and Sheinberg, 1996; Tanaka, 1996). Evidence for gradual increases in tolerance for changes in position and scale is indirectly suggested by the presence of both simple and complex cells in V1 (Hubel and Wiesel, 1965) as well as increases in receptive field (RF) size along the ventral stream (Kobatake and Tanaka, 1994).

At the same time, many open questions remain. First, although highly selective IT neurons do exist, most neurons in IT are broadly tuned for different objects when tested with large sets of images (Desimone et al., 1984; Rolls and Tovee, 1995; Kreiman et al., 2006; Zoccolan et al., 2007). Thus, it remains unclear whether selectivity is increasing across the pathway "on average." Second, remarkably few studies have directly compared different ventral stream visual areas using the same stimuli under the same conditions, and, when direct comparisons are made, they fail to find clear distinctions between areas (Hegdé and Van Essen, 2007). Notably, direct and definitive comparisons are difficult to make given our lack of understanding of the visual features that activate neurons beyond V1. Third, recent computational work demonstrates that the invariant object recognition problem could be solved by a distributed representation across a population of neurons with small receptive fields (Li et al., 2009) and hence earlier in the pathway then previously appreciated. Finally,

Received Jan. 12, 2010; revised July 23, 2010; accepted Aug. 4, 2010.

This work was funded by National Eye Institute Grants 1F32EY018063 and R01EY014970 and the McKnight Endowment Fund for Neuroscience. We thank Nuo Li, John Maunsell, Tomaso Poggio, Eero Simoncelli, and Davide Zoccolan for helpful discussions. We also thank Ben Andken, Jennie Deutsch, Marie Maloof, and Robert Marini for technical support.

Correspondence should be addressed to Nicole Rust, Department of Psychology, University of Pennsylvania, 3401 Walnut Street, Room 317C, Philadelphia, PA 19104. E-mail: nrust@sas.upenn.edu.

DOI:10.1523/JNEUROSCI.0179-10.2010

 0270-6474/10/3012978-18$15.00/0

the single-neuron response properties supporting these putative increases in selectivity and invariance remain little understood.

Given that selectivity and tolerance are thought to lie at the crux of the object recognition problem, we aimed to directly examine whether and how they change along the ventral stream and thus lay the groundwork for additional quantification of these computations so as to meaningfully constrain computational models. Our results show that the visual representation is reformatted between two stages of the ventral stream: at both the population level and the single-unit level, we document an increase in selectivity for naturally occurring conjunctions of simple visual features and an increase in tolerance to identity preserving transformations.

Materials and Methods

Animals and surgery

Experiments were performed on two male rhesus macaque monkeys (*Macaca mulatta*) weighing 5.0 and 8.0 kg. Aseptic surgery was performed to implant a head post and scleral search coil in each animal before the onset of training. An additional one to two surgeries were performed to place recording chambers over both hemispheres of V4 and IT. All surgical and animal procedures were performed in accordance with the National Institute of Health guidelines and the Massachusetts Institute of Technology Committee on Animal Care.

Stimuli and task

All behavioral training and testing was performed using standard operant conditioning (juice reward), head stabilization, and high-accuracy, real-time eye tracking. Stimuli, reward, and data acquisition were controlled using customized software. Stimuli were presented on a cathode ray tube monitor with an 85 Hz refresh rate positioned 49 cm away such that it subtended 44 × 33°. All images were presented at the center of gaze, in a circular aperture that blended into a gray background (see Figs. 1*a*, 2). Both monkeys were trained to initiate each trial by fixating a central red point (0.15°) within a square fixation window that ranged from ±0.9° to ±1.1° for up to 4 s. Across the repeated presentations of a stimulus recorded from a neuron, deviation of the eye position (measured relative to the mean position across all trials) was extremely small: on average, 82, 86, and 97% of presentations occurred within windows with a radius of 0.05, 0.1, and 0.25°, respectively. Soon after initiating fixation (250 ms), a series of visual stimuli were presented in rapid succession (each for 218 ms or approximately five per second) with no intervening blank period. This presentation duration is consistent with that produced spontaneously by a free-viewing monkey (DiCarlo and Maunsell, 2000) and is sufficient for successful object discrimination (see below). Monkey 1 was rewarded with juice for maintaining fixation for 2.43 s (10 stimuli). Monkey 2 viewed the same images while engaged in an invariant object detection task that required a saccade to a response dot 10° below the fixation point after encountering an image that contained a motorcycle (see Fig. 2) (supplemental Fig. 3, available at www.jneurosci.org as supplemental material) to receive a reward. After onset of the motorcycle stimulus, the animal had 500 ms to reach the response window; after the 218 ms presentation of the motorcycle image ceased, other images continued to stream. The motorcycle image was presented as the *N*th image, where *N* was randomly selected from a uniform distribution ranging from 2 to 20. To fully engage the system involved in invariant object recognition, the same motorcycle was presented at different positions, scales, and on different backgrounds. Each day, the monkey viewed the same 28 motorcycle photographs and an additional two novel motorcycle images (supplemental Fig. 3, available at www.jneurosci.org as supplemental material). Performance on this task was high (miss rate for familiar motorcycle images of 1.32%, range of 0.48–2.4%; miss rate for novel motorcycle images of 1.37%; false alarm rate of 11%; mean reaction time of 254 ms).

Images. Designed to probe V4 and IT selectivity and tolerance, the image set included 155 images (supplemental Figs. 1, 2, available at www.jneurosci.org as supplemental material). Included were 50 natural images and 50 scrambled versions of those images (see Fig. 4) (scrambling procedure described in detail below). For 10 of the natural images, five additional transformations were also presented (rescaled to 1.5× and 0.5×; shifted 1.5° left and right; and presentation in the context of a natural background) (see Fig. 6). An additional five blank (gray) stimuli were included to measure the baseline firing rate. Ten repeats of each stimulus were collected.

At the onset of each trial, one of the images was randomly presented, and the responses to this stimulus were disregarded to minimize onset transient effects. Thereafter, stimuli were presented in a random sequence, and each stimulus was presented once before re-randomization. If the monkey's eyes moved outside the fixation window, the trial was immediately aborted and the remaining stimuli on that trial were included in a re-randomization with the remaining images. Both monkeys were exposed to the image set for at least 3 weeks before data collection.

Image scrambling. Images were scrambled using a texture synthesis method introduced by Portilla and Simoncelli (2000), using a publicly available Matlab (MathWorks) implementation (http://www.cns.nyu.edu/~lcv/texture/). Briefly, this method extracts 710 parameters from an original image and then generates a scrambled image by repeatedly forcing a new image (initially filled with Gaussian white noise) to match these parameters. The parameters are all obtained by averaging local measurements across all spatial positions within the original image and are thus altogether capable of representing the number and type of local features in the original image, while lacking information about their specific locations. Some parameters describe pixel statistics (mean, variance, skew, and kurtosis) of the image, ensuring that the synthesized image is matched in luminance distribution. Most parameters describe averages of various local combinations of oriented linear filter responses. Filter responses are computed using a complex-valued steerable pyramid decomposition, which approximates the response of a population of model V1 complex cells tuned for different orientations and scales (in our case, four orientations and four scales that included filters one-eighth, one-quarter, one-half, and one times the size of the image or equivalently 0.625, 1.25, 2.5 and 5°) and that tile all positions in the image. The parameters include the local autocorrelation of the linear filter responses, which enables the representation of periodic structures. Also included are the correlations of complex magnitudes of nearby pairs of filters tuned for the same scale, and the same or different orientations, which enables the representation of rudimentary oriented feature information (e.g., lines, edges, corners, and junctions). To capture the alignment of phase structure in local features, the model also includes a form of cross-scale phase correlation. Finally, several moments of the residual low- and high-frequency nonoriented sub-bands are retained; the combination of these parameters with those described above captures the spatial frequency (spectral) content. The synthesis algorithm is not guaranteed to converge to a solution with matched parameters, and we discarded the occasional scrambled image that failed to match in either luminance or spatial frequency content. On rare occasions, we also discarded scrambled images that (by chance) retained some global structure and (by visual inspection) appeared to contain an identifiable object. To quantify the degree of convergence for the 50 images included in this study, we began by normalizing all parameters to have the same mean (0) and variance (1) across all 50 natural images and their 50 scrambled counterparts. We then computed the Pearson's correlation coefficient between the normalized parameters for pairs of images. The correlation coefficients for natural images and their corresponding scrambled image pairs were very close to perfect (mean of 0.9926, range of 0.9707–0.9995). For comparison, the parameter similarity of other possible image pairings in our image set (i.e., all noncorresponding natural and scrambled image pairings) ranged from −0.7432 to 0.8939. This confirms that the algorithm successfully converged for the images we included in this study and that the scrambled version of each image is much closer to its natural counterpart than to all other images in the set [when measured in a Portilla and Simoncelli (2000) image basis].

Receptive field mapping (V4). Designed to measure the location and extent of V4 receptive fields, bars were presented, each for 500 ms, one per trial, centered on a 5 × 5 invisible grid. In an additional baseline condition, the fixation point was presented without a bar stimulus. Bar orientation, polarity (black or white), length, and width as well as the grid

center and extent were adjusted for each cell based on preliminary hand mapping. On each trial, the monkey was required to maintain fixation on a small response dot (0.125°) to receive a reward. The responses to three repeats were collected at each position.

Recording procedures

The activity of well-isolated V4 and IT neurons was monitored serially using standard single microelectrode methods (Zoccolan et al., 2005). Recorded signals were amplified, filtered, and fed into a time–amplitude window discriminator. Electrodes used to record from V4 and IT were constructed from the same materials (glass-coated tungsten) by the same manufacturer (Alpha Omega) and matched in impedance ($\sim$0.5 MΩ).

Before each recording session, an electrode was advanced to the appropriate visual area. After allowing the electrode 15–30 min to settle, it was then slowly advanced through the cortex until a waveform was isolated. Great care was taken to ensure that any neuron whose waveform could be isolated would be recorded, regardless of baseline or visually elicited firing rate. In cases in which the electrode traversed the gray matter approximately perpendicularly (the lower visual field representation of V4 and all penetrations of IT), care was taken to ensure that all layers were sampled approximately uniformly. While searching for cells, the monkey engaged in the same task required during the data collection (described above). This included periods of viewing stimuli interleaved with intertrial epochs in which no stimuli were presented and the monkey was free to look around the room. Additionally, no data analysis was performed during data acquisition to assess the "quality" of the neuron; all neurons were recorded until the experiment was complete or until the waveform was lost.

To guard against nonstationary effects (e.g., familiarity with the images), recordings in each animal were alternated between V4 and IT. Specifically, recordings were made in one visual area (V4 or IT) for 1–5 weeks, and then recordings were made in the other area; this alternating process was repeated until all data were collected.

The left and right hemispheres of both V4 and IT were recorded in each monkey (four recording chambers for each subject, resulting in eight chambers in total). Both hemispheres of each visual area were sampled approximately equally in each monkey, with approximately twice as many cells sampled in monkey 2 compared with monkey 1 (monkey 1: V4 left, $n = 25$; V4 right, $n = 23$; IT left, $n = 32$; IT right, $n = 16$; monkey 2: V4 left, $n = 42$; V4 right, $n = 50$; IT left, $n = 35$; IT right, $n = 60$). Chamber placements varied slightly between hemispheres and between animals and were guided by anatomical magnetic resonance images. A representative IT chamber was centered 15.5 mm anterior of the ear canal, 12 mm lateral of the midline and angled 5° lateral. The resulting region of IT recorded was located on the ventral surface of the brain, lateral to the anterior middle temporal sulcus and spanned $\sim$10.5–17.5 mm anterior to the ear canals (Felleman and Van Essen, 1991). A representative V4 chamber was centered 6 mm posterior and 17 mm dorsal to the ear canals. V4 recording sites were confirmed by a combination of receptive field size and location (see Fig. 1*b–d*). V4 receptive fields in lower visual field, found between the lunate and superior temporal sulcus (STS), were confirmed as having receptive field centers that transversed from the vertical to horizontal meridian across posterior to anterior recording locations as well as receptive field sizes as a function of eccentricity that were consistent with results reported previously (Desimone and Schein, 1987; Gattass et al., 1988). Neurons with receptive fields at the fovea and near the upper visual field were more difficult to verify given their existence within the inferior occipital sulcus (IOS) and at the foveal confluence of V1, V2, and V4. Thus, it is not certain that all the neurons in the upper field were from V4, although the receptive field sizes were more consistent with V4 than either V2 or V1. Notably, given the absence of easily identifiable boundaries in this region, anatomical reconstruction would not assist in verifying their precise location. We also note that, aside from their receptive field locations, neurons in the upper visual field did not have any obvious, distinguishable properties from those in the lower visual field. Moreover, the claims of this study (a comparison between mid-level and high-level visual areas) would be little affected by the occasional corruption of a neuron from a nearby visual area.

Analysis

Spike sorting. Spike waveforms were isolated online using a dual window discriminator. In addition, a *post hoc*, template-based spike-sorting procedure was applied to remove spurious electrical artifacts and corruption by other neurons. Specifically, from the data collected from all trials of an experiment from a single neuron, we calculated the mean and SD of all collected waveforms. We began by setting the boundary criteria for an accepted waveform as the mean $\pm$ 2 SDs and then adjusted the criteria, by eye, to maximize the inclusion of spikes with the same waveform shapes while minimizing the inclusion of spikes with different waveforms shapes. In all cases, $<$10% of electrical events were disregarded, and, for all the results reported here, qualitatively similar results were obtained with the raw and spike-sorted data.

Latency. We computed the responses of each neuron by counting spikes in a window matched to the duration of the stimulus (218 ms) and shifted to account for the latency of the neuron. To calculate the latency of each neuron, we used techniques similar to those described by Zoccolan et al. (2007). Briefly, we began by finding the stimuli that evoked at least 70% of the peak response and then used these stimuli to calculate the latency of the neuron. Specifically, we binned neuronal responses in a 218 ms window with a latency guess of 75 ms for V4 and 110 ms for IT, and, using these values, we computed the mean firing rates to each stimulus for a given experiment. We then computed a combined peristimulus time histogram across all stimuli that evoked $>$70% of the peak response with 10 ms overlapping bins shifted in time steps of 1 ms. Background firing rate was estimated as the mean firing rate 100 ms before stimulus onset, and latency was estimated as the first bin that exceeded 15% of the peak firing rate relative to the background rate. All latency estimates were examined by eye and, when required, adjusted. We used the same, mean latency for all the neurons in a given visual area (V4, 78 ms; IT, 113 ms). Using either a fixed latency for all neurons or tailoring the latency for each neuron resulted in qualitatively similar results across the population. Additionally, we performed each analysis at different latencies and found that the results we report here were robust over a wide range of latency offsets (data not shown).

V4 receptive field mapping. Only those neurons that produced clear visually evoked responses (see above) at a minimum of one position were considered for receptive field position analysis. The center of the receptive field was estimated by fitting a two-dimensional, oriented Gaussian to these data (Op De Beeck and Vogels, 2000) and confirmed by visual inspection. Simulations of two-dimensional receptive field profiles with Poisson spiking variability confirm that measuring RF position in this manner produces a robust estimate of the receptive field center (although receptive field size, not reported here, may require more than three repeated trials for some neurons).

Population discriminability. To determine how well a population of neurons could discriminate between a set of images, we implemented a linear classifier readout procedure similar to that used by Hung et al. (2005). Starting with the spike count responses of a population of N neurons to P presentations of M images, each presentation of an image resulted in a population response vector $\mathbf{x}$ with a dimensionality equivalent to $N \times 1$ (see Fig. 3, left), where repeated presentations of the same images can be envisioned as a cloud in an N dimensional space (see Fig. 3, middle). The linear readout amounted to finding a linear hyperplane that would best separate the response cloud corresponding to each image from the response clouds corresponding to all other images (see Fig. 3, lines). More specifically, the linear readout took the following form:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b,$$

where $\mathbf{w}$ is a $N \times 1$ vector describing the linear weight applied to each neuron (and thus defines the orientation of the hyperplane), and b is a scalar value that offsets the hyperplane from the origin and acts as a threshold. We used a standard "one-versus-rest" training and testing classification scheme (Hung et al., 2005; Li et al., 2009). Specifically, one such linear classifier was determined for each image or grouped set of images (see details below). To determine the population "decision" about which image (or object) was presented, a response vector $\mathbf{x}$, corresponding to the population response of one image, was then applied to

each of the classifiers, and the classifier with the largest output [the classifier with the largest, positive $f(\mathbf{x})$] was taken as the choice of the population (see Fig. 3, right). Performance was measured as the proportion of correct answers when asked to identify each image with independent data not used for training (i.e., standard cross-validation), and details of the cross-validation data cuts are below.

The hyperplane and threshold for each classifier was determined by a support vector machine (SVM) procedure using the LIBSVM library (http://www.csie.ntu.edu.tw/~cjlin/libsvm) with a linear kernel, the C-SVC algorithm, and cost (C) set to 0.5. To avoid overfitting, we always used a cross-validation procedure to assess performance. To test scrambling sensitivity (see Figs. 5, 9, 10*a*), 80% (8 of 10) of the trials for each image were used for training the hyperplanes, whereas 20% (2 of 10) of the trials were set aside for cross-validated testing. To test generalization (see Fig. 7), 80% of the trials corresponding to the reference image were used to train the hyperplanes, and 20% (2 of 10) of the trials at each transformation were used to assess cross-validated performance. To test linear separability (see Figs. 8, 11, 12*b*,*d*, 13*a*), 80% of the trials for each transformation were combined and used to train the hyperplanes, and performance was assessed with the remaining 20% of trials. To equalize firing rates across neurons, each neuron was normalized to have the same (zero) mean and (unit) SD across M stimuli before training and testing. When comparing performance across two conditions (e.g., the natural vs scrambled images in Fig. 5 or when probing generalization in Fig. 7), the responses of neurons were normalized across the responses to all images sets combined. For Figures 5, 7, and 8, qualitatively similar results were produced when using raw (non-normalized) spike counts.

In many plots, we report performance as a function of the number of neurons (randomly selected) included in the analysis. To measure the variability that can be attributed to the particular subpopulation of neurons selected as well as the particular trials used for training and testing, we applied a resampling procedure. On each iteration of the resampling, a new subpopulation of neurons were randomly selected (without replacement) from all neurons, and trials were randomly assigned for training and testing (without replacement). Error bars were calculated as the SD of performance across 50 iterations. We also computed chance performance by randomly assigning the images associated with each response vector and then performing the analysis as described above.

To measure the ability of a population to generalize across an object-identity-preserving image transformation (e.g., a change in the position of an object), the hyperplanes were first trained and tested on a "reference" condition. Generalization was measured as identification performance when this representation was tested with the responses to the transformed images (see Fig. 7). Generalization capacity (see Fig. 7*d*) was calculated as the ratio of the mean performance in the generalization condition and the performance in the reference condition. To calculate confidence intervals on this metric, generalization capacity was calculated on each iteration of a resampling procedure in which trials were randomly assigned for training and testing without replacement. Error bars were calculated as the SD of performance across 50 iterations. To more specifically test the linear separability of the V4 and IT representations over all the object-identity-preserving image transformations (see Fig. 8*a*), we performed an additional analysis in which hyperplanes were trained with 80% of the data to simultaneously group all six transformations of each object (see Fig. 6), and the representation was tested with 20% of randomly selected trials not included in the training. Error bars were calculated as the SD of performance across 50 iterations (see Fig. 8*b*).

To investigate how robust our results were to the particular readout technique, we assessed population performance using two additional readout methods. The first was a correlation-based classifier (Meyers et al., 2008). Here, a classifier for each object was determined as the mean response vector (see Fig. 3, left) across all trials of the training data. The response of each classifier to a (separately measured) "test" response vector was determined as the Pearson's correlation coefficient between the classifier and the test; the classifier with the largest positive coefficient was taken as the "decision" of the population. The correlation-based classifier is simpler than the SVM in that (1) the effective "weights" placed on each neuron are determined by the normalized firing rates of the neuron as opposed to an optimization procedure and (2) the threshold is always set to zero. Sampling subpopulations of neurons, assigning data for training and testing and calculation of error was identical to that described for the SVM linear classifier.

As a third measure of population discriminability, we measured the normalized Euclidean distance between the response clouds (see Fig. 3) for the natural and scrambled images in both V4 and IT. Before calculating this population measure, the response for each neuron was normalized to have a maximal average firing rate (across all trials) of 1 over all the tested images. Given a set of T response vectors $\{\mathbf{x}_i\}$ that describe the population response on T trials to one stimulus and a second set of vectors $\{\mathbf{y}_i\}$ that describe the response on T trials to a second stimulus, normalized Euclidean distance was calculated as the mean distance between the mean of one response cloud ($\bar{\mathbf{x}}$) and all other trials of a second response cloud ($\mathbf{y}_i$), normalized by the SD along each dimension of $\mathbf{x}_i$, σ_x:

$$d(x,y) = \frac{\sum_{i}^{T} \left\| \frac{\bar{\mathbf{x}} - \mathbf{y}_i}{\sigma_x} \right\|}{T}.$$

For a small fraction of neurons, a small fraction of images (<4% in total) failed to evoke a response on any trial, resulting in $\sigma_x = 0$. If a neuron failed to spike in response to all of the repeated presentations of an image, that neuron was disregarded from the distance measures for that image. The distance between two image sets was calculated as the geometric mean of the normalized Euclidean distance across all possible pairwise distance measures for each set of 50 images ($N = 50^2 - 50 = 2450$).

Single-neuron received operating characteristic analysis. We quantified the degree of overlap between two spike count distributions with a receiver operating characteristic (ROC) analysis (Green and Swets, 1966; Britten et al., 1992). Given two spike count distributions that arise from two different alternatives (e.g., one image vs another image for Fig. 9*a*, or all the transformations of one object vs all the transformations of all other objects for Fig. 11), we generated an ROC curve by computing the proportion of trials for alternative 1 on which the response exceed the criterion versus the proportion of trials for alternative 2 on which the response exceeded the criterion for a range of criteria. The ROC value was taken as the area under this curve.

Single-neuron measurement of linear separability. To obtain a single-unit measure that is indicative of population linear separability for objects across the different transformations described in Figure 6, we applied a metric described previously (Brincat and Connor, 2004; Janssen et al., 2008) and reviewed below. We here term this metric single-neuron "linear separability index" (se Fig. 12*c*) because, although it does not indicate linear separability of object identity in the response of a single neuron, under reasonable assumptions about the distribution of tuning functions in a population, it is a good predictor of linear separability of object identity at the neuronal population level and is a far better predictor than single-unit measures such as receptive field size (Li et al., 2009). This metric measures how well the response of a neuron can be explained as resulting from independent tuning for object identity and the identity-preserving image transformation (e.g., retinal position, size). We begin by splitting the neuron's response data into two halves by randomly assigning the 10 repeated trials of each stimulus into two sets of five trials. From these data, we compute two matrices (M_1 and M_2), each described by the mean firing rate response to each of 10 objects across the six transformations described in Figure 6 (i.e., each M was 10×6). Because an artificially high linear separability index can result from the degenerate case in which a neuron responds to objects under only one of the transformed conditions (e.g., only when the objects were positioned to the left), we only included neurons that responded significantly differently than baseline (to any object) under at least two of the transformed conditions (two-tailed t test, $p < 0.05$). For these neurons, we computed the independent tuning prediction by computing the singular value decomposition of one matrix ($M_1 = USV'$) and then used the first principal

component to determine the expected response to each object across the six transformations (M_{pred} was the product of the first columns of U and V′). Finally, we computed the element-wise Pearson's correlation of M_{pred} and the separately measured M_2. The resulting metric was bounded at −1 to 1.

Alignment of linear separability in simulation. To increase the average single-neuron linear separability index of the V4 population to match the average in IT (see Fig. 13*c*), we began by computing the singular value decomposition of 10 object × 6 response surface of each V4 neuron (M_1 = USV′). We then reconstructed a new response surface (M_{new}) with a higher single-neuron linear separability index by adjusting the weights of the diagonal matrix S. Specifically, we fixed the weight of the first (separable) component to its original value and rescaled the remaining entries of S by the same multiplicative factor (<1) whose value was chosen to match the mean linear separability in the simulated V4 population and the mean of the IT population. The modified response surface was computed as (M_{new} = US_{new}V′).

Results

We took a population-based, comparative approach toward determining how selectivity and tolerance change along the ventral visual pathway. Specifically, we compared neurons in mid-level visual area V4 with the last stage of purely visual processing, IT. Comparing the response properties of neurons across different visual areas can be difficult as a result of differences in the sizes of their receptive fields: although it makes little sense to measure the response properties of a neuron outside its receptive field, neurons at earlier stages of the ventral pathway typically have receptive fields that are smaller than the objects they encounter, and rescaling the objects to fit entirely within these small receptive fields does not make sense if one's goal is to study the representation of real-world images along the ventral stream. Thus, in contrast to traditional single-neuron approaches, we started with the assumption that comparisons between V4 and IT would be much more insightful if we evaluated the combined behavior of neurons in each area as a population and with respect to the kinds of visual tasks the ventral stream is likely to support. In our experiments, stimuli were always presented in a fixed retinal location and at a fixed retinal size despite the receptive field location of the neuron we were recording. Specifically, all stimuli were presented in a 5° diameter circular aperture placed at the center of gaze (Fig. 1*a*). Neurons in IT have receptive fields that often encompass the entire 5° image; these receptive fields typically include the center of gaze and extend into all four visual quadrants (Fig. 1*a*, right) (Op De Beeck and Vogels, 2000). The organization of V4 is quite different: V4 receptive fields are retinotopically organized and are primarily confined to the contralateral hemifield (Fig. 1*a*, left) (Desimone and Schein, 1987; Gattass et al., 1988). To compare the representation of images in these two visual areas, we recorded from V4 neurons whose receptive fields tiled the image (Fig. 1*b*) and compared the V4 population responses to a similarly sized population of IT cells. This required us to record from both hemispheres of V4 and within each hemisphere to sample neurons with receptive fields in both the upper and lower visual quadrants (Fig. 1*c*,*d*). Similarly, because IT receptive field centers tend to be shifted toward the contralateral visual field (Op De Beeck and Vogels, 2000), we also recorded from both hemispheres of IT. While we were recording, one monkey performed a challenging object detection task to engage the ventral visual stream and maintain a constant level of arousal (Fig. 2) (supplemental Fig. 3, available at www.jneurosci.org as supplemental material), whereas a second monkey was passively viewing the images while fixating. We found no differences in the effects observed between the two monkeys (see Tables 1, 2), and, unless noted, the data presented here are pooled across both subjects.

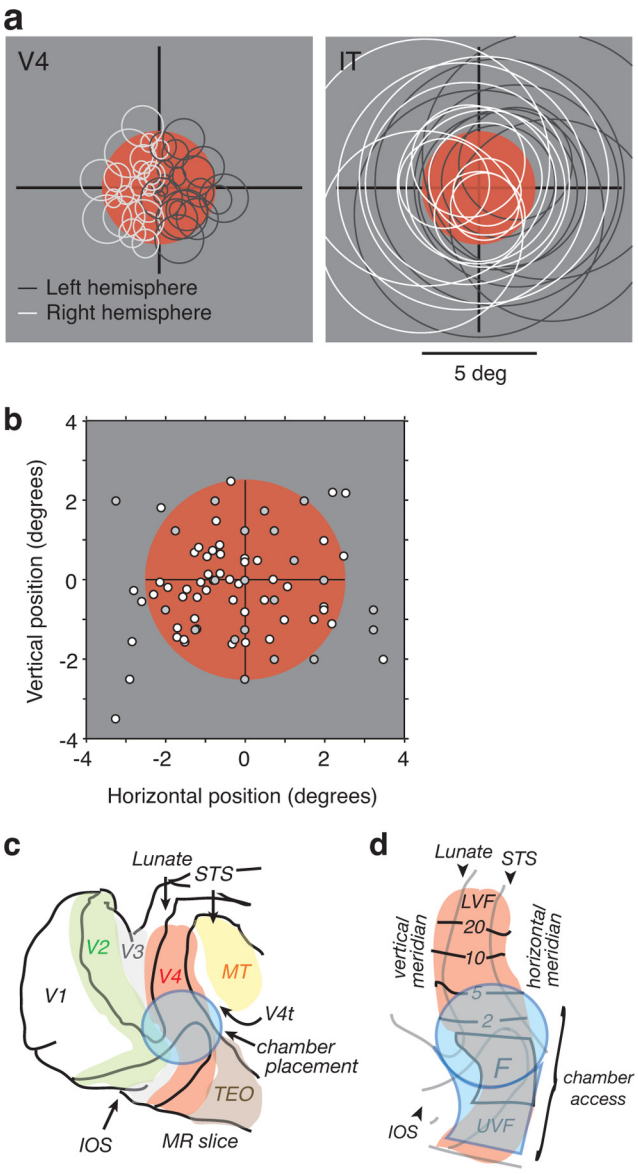

Figure 1. Experimental design. ***a***, All images were displayed in a 5° diameter aperture located at the center of gaze (red). Expected receptive field locations and sizes for neurons in V4 (Desimone and Schein, 1987; Gattass et al., 1988) and IT (Op De Beeck and Vogels, 2000). To compare these two areas, we targeted V4 neurons such that the population of V4 receptive fields tiled the image. This required recording from both the right (white) and left (dark gray) hemispheres. ***b***, The receptive field locations of a subset (78 of 140) of V4 neurons recorded; dots illustrate their centers relative to the 5° diameter stimulus aperture (gray, monkey 1; white, monkey 2). ***c***, Occipital cortex, illustrating the location of V4 relative to other visual areas, adapted from Gattass et al. (1988). V4 exists on the cortical surface between the lunate sulcus and the STS and extends into the IOS. Approximate chamber placement indicated in cyan. ***d***, Expanded view of V4, also adapted from Gattass et al. (1988). The lower visual field representation (LVF) in V4 exists on the cortical surface, in which receptive field locations move toward the fovea as one traverses ventrally; approximate eccentricities are labeled according to Gattass et al. (1988). At all eccentricities, receptive fields cluster toward the vertical meridian near the lunate and move toward the horizontal meridian as one approaches the STS. The foveal representation, (labeled F) begins at the tip of the IOS. The upper visual field representation (UVF) can be found within the IOS. Given the foveal confluence of V1, V2, and V4 within the IOS, it is not certain that all of the neurons in the upper field were in fact from V4, although the receptive field sizes were more consistent with V4 than either V2 or V1. Cyan illustrates the approximate region that can be accessed via the chamber, which includes both the lower and upper visual field representations. MT, Middle temporal area; TEO, temporal–occipital area.

To compare the representation of images between the V4 and IT populations, we performed a variety of analyses that all sought

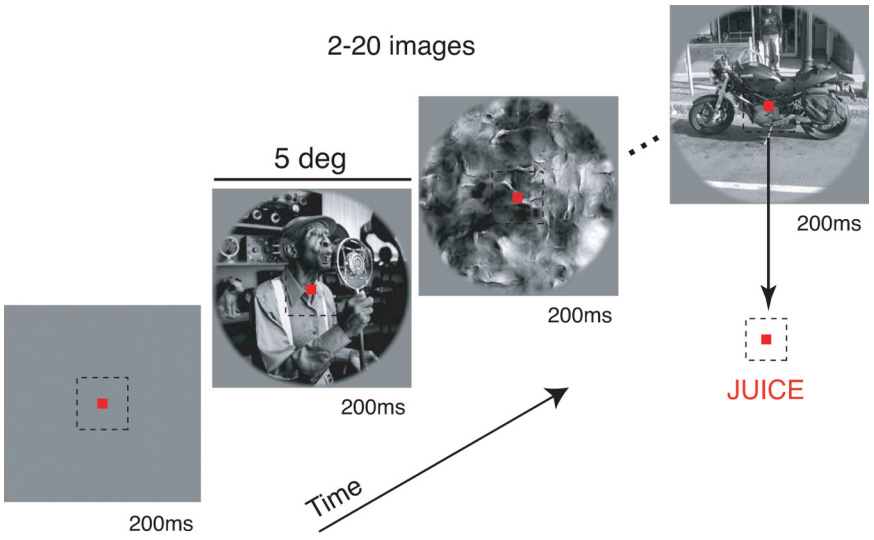

Figure 2. The object detection task performed by monkey 2. Each trial began with the monkey looking at a fixation point. After a brief delay, images were presented, in random order, each for 200 ms. At a randomly preselected point in the trial, an image containing a motorcycle appeared. The monkey then had 500 ms to saccade to the response dot to receive a juice reward. In the intervening time, images continued to stream. To ensure that the monkey was performing an object recognition task as opposed to relying on low-level visual cues, the motorcycle was presented at different scales, positions, and on different backgrounds. In addition, novel motorcycle images were introduced each day (supplemental Fig. 3, available at www.jneurosci.org as supplemental material).

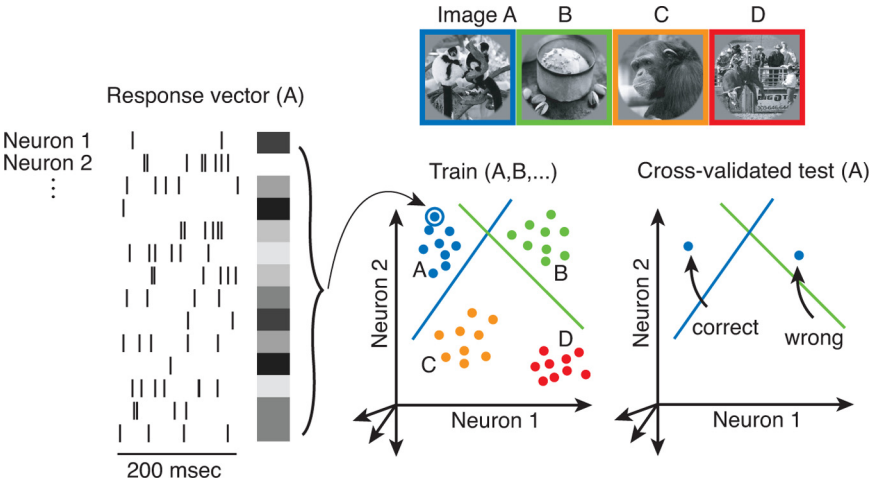

Figure 3. Assessing the ability of a population of neurons to encode an image set by measuring discriminability with a linear population readout. Left, A hypothetical population response for a single presentation of an image (labeled A). After adjusting for latency (see Materials and Methods), spikes were counted in a 200 ms window. The spike counts for the *N* neurons recorded within a given visual area were combined to form a "response vector" of length *N*. Right, The response vector exists in an *N*-dimensional space but is illustrated in the two-dimensional space defined by the responses of neurons 1 and 2 (circled blue dot). Because neurons are noisy, different presentations of the same image produce slightly different response vectors and together all presentations form a "response cloud." The images producing each response vector are labeled by color. The ability of the population to discriminate between different images is proportional to how far apart the response clouds are in this space. We quantified discriminability using linear classifier readout techniques (see Materials and Methods). This amounted to finding, for each image, the optimal linear hyperplane (shown here as a line) that separated all the responses to that image from all the responses to all other images. After using a subset of the trials to find each hyperplane, we tested discriminability with other trials by looking to see where the response vectors fell. The hyperplane that produced the maximal response (the hyperplane for which the response vector was on the correct side and the farthest from the boundary) was scored as the answer, and performance was measured as the percentage correct on this image identification task. Example correct and wrong answers for presentations of stimulus A are shown (right).

to measure how well the combined activity of each population could discriminate between different images within a set (Hung et al., 2005; Li et al., 2009). To quantify discriminability, we trained each population using a linear readout scheme (represented by the lines in Fig. 3, middle) and then tested the representation on an image identification task (Fig. 3, right) (see Materials and Methods). Linear readout rules are neurally plausible in that they are equivalent to a neuron at the next level of processing that receives weighted excitatory and inhibitory input from the population in question, followed by a threshold. It is important to recognize that, by using a linear readout, we are probing the information that is explicitly accessible at a given level of visual processing compared with "total information," which also includes information that is present but not accessible using simple neural machinery. Assuming that the information that propagates to IT does so by passing through V4, the quantity of interest is not "total" but "accessible" information at each stage. Notably, measures of performance on these discriminability tasks depend not only on the format of the population representation but also on the number of neurons included in the analysis and the number of images included in a set. Thus, although absolute performance on any one analysis may be difficult to interpret, this type of analysis is useful for making comparisons, either between stimulus classes or between visual areas.

Comparison of selectivity in V4 and IT

We begin by describing our measurements of selectivity in V4 and IT. Here, we will be measuring a form of selectivity that describes the complexity of image features that activate a neuron, which we refer to as "conjunction sensitivity." Specifically, V1 neurons are known to be activated by small image patches that contain energy at a particular orientation and spatial frequency (for review, see Lennie and Movshon, 2005); a hypothetical neuron that requires a more complex conjunction of local oriented segments to evoke a response would have a higher conjunction sensitivity than a neuron in V1. We aimed to test the hypothesis that IT neurons require a more complex conjunction of features (have a higher conjunction sensitivity) than V4 neurons. Traditionally, this hypothesis has been difficult to systematically test given the lack of understanding of the types of image features that activate V4 and IT neurons. For example, attempts have been made to find the simplest "critical features" that drive neurons in different visual areas and then compare their complexity (Kobatake and Tanaka, 1994), but given that such techniques involve specifically tailoring the stimulus set for each neuron, these methods produce results that are difficult to systematically compare across different neurons and across different areas. Because we wanted to

systematically compare visual areas under matched conditions that included the same stimuli presented under the same conditions, we attempted a different approach in which we measured the sensitivity of populations of neurons to image scrambling ("scrambling sensitivity") as an estimate of the conjunction sensitivity for that population. After describing the image scrambling procedure itself, we describe the rationale behind these experiments. This procedure for scrambling, introduced by Portilla and Simoncelli (2000), minimizes the artifacts common to many scrambling procedures (such as introduction of high spatial frequencies) and, moreover, preserves the types of image features thought to be encoded by the first visual area in the cortex, V1. Specifically, given a natural image, this procedure produces a second image containing the same number and type of local, oriented elements but presented at random positions within the image (Fig. 4) (see Materials and Methods). Viewing these images shows that they are clearly rich in local structure, but, because the positions of the local features have been randomized, they contain no definable objects. One important property of V1 neurons that is reproduced by the model is the covariation between the size of the elements and their spatial frequency content (Ringach, 2002): larger elements contain lower spatial frequency content (e.g., large, oriented blobs), whereas high spatial frequency content is encoded by smaller elements. Consequently, because long, continuous lines are produced by higher spatial frequencies, they are broken up into a number of local elements and tend to be destroyed in the scrambled images. Similarly, contours, shape, figure/background, and at least some of the global structure often referred to as "gist" are destroyed by this scrambling procedure. Our operational definition of "feature conjunctions" for the purposes of this study is the destruction of these higher-order properties of images, which remain statistically ill-defined, after using the scrambling procedure described above. Consideration of how our results depend on the particular scales at which we scrambled the images can be found in Discussion.

The rationale behind comparing discriminability among natural images with the discriminability among scrambled images is as follows. In a first scenario, a population of neurons that are tuned to the local features in an image (that is, they have, by definition, little or no conjunction sensitivity) should encode both intact natural images and appropriately scrambled versions of those images with similar fidelity. That is, if the scrambling process preserves the types of local features for which single neurons are tuned but merely rearranges them, scrambling will change the particular pattern of neurons activated but will not affect the ability of the population to encode (i.e., discriminate among) the scrambled images (Fig. 5*a*, left). In a second, alternative scenario, a population of neurons may only be activated by the specific, feature conjunctions found in natural images, and, because scrambling destroys these conjunctions, the neurons would fail to respond to (and thus fail to discriminate among) such images (Fig. 5*a*, right). A third scenario is also possible: a population may be tuned for random conjunctions of local features and therefore have no preference for naturally occurring feature configurations. Similar to the first scenario, such a population would discriminate among both natural and scrambled images with similar fidelity. In other words, matched population discriminability for natural and scrambled images is consistent both with a population that encodes local structure and with a population that encodes random conjunctions of features. However, reduced discriminability for scrambled compared with natural images indicates a population that preferentially encodes the feature conjunctions found in natural images.

We begin by comparing the ability of the V4 versus the IT populations to discriminate among different members of a set of

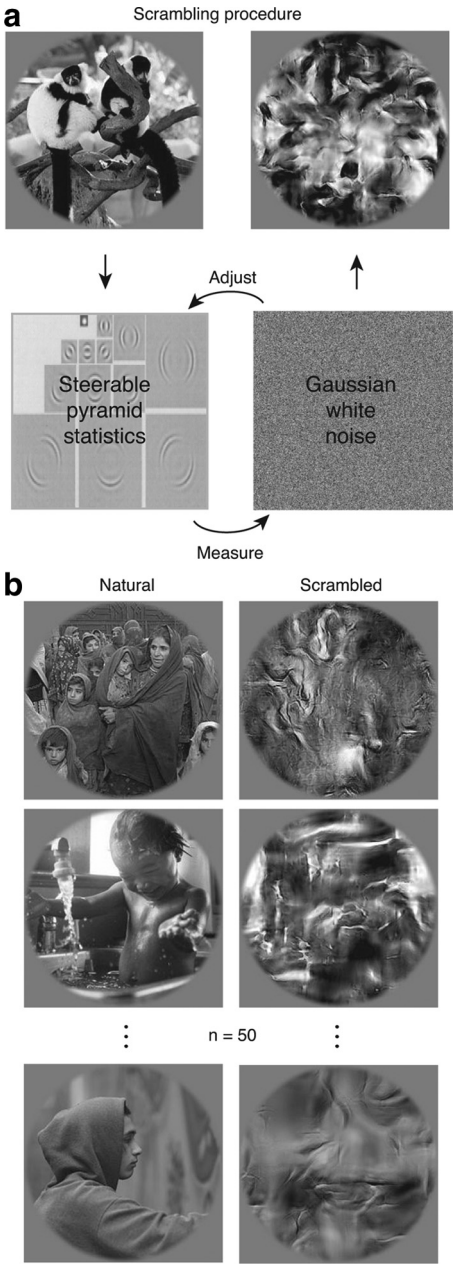

Figure 4. Scrambling procedure. ***a***, Images were scrambled using a model introduced by Portilla and Simoncelli (2000). Briefly, the procedure begins by computing a number of image statistics. A Gaussian white-noise image is then iteratively adjusted until it has the same number and type of local, oriented elements but presented at random positions within the image (see Materials and Methods). ***b***, Example natural images and their scrambled counterparts. Each set contained 50 images.

50 natural images using linear readout rules (specifically, SVMs; see Materials and Methods). We found that encoding performance increased as a function of the number of neurons included in the analysis, and performance was similar but slightly higher in V4 compared with IT (Fig. 5*b*,*c*, black). This nearly equivalent population discriminability in V4 and IT for natural images with a comparable number of neurons (sampled without bias; see Materials and Methods) has not, to our knowledge, been described previously, and it was not a predetermined result of our approach; this result implies that information about natural images is primarily maintained as signals travel from V4 to IT (assuming an approximately equal number of neurons in V4 and IT). Crit-

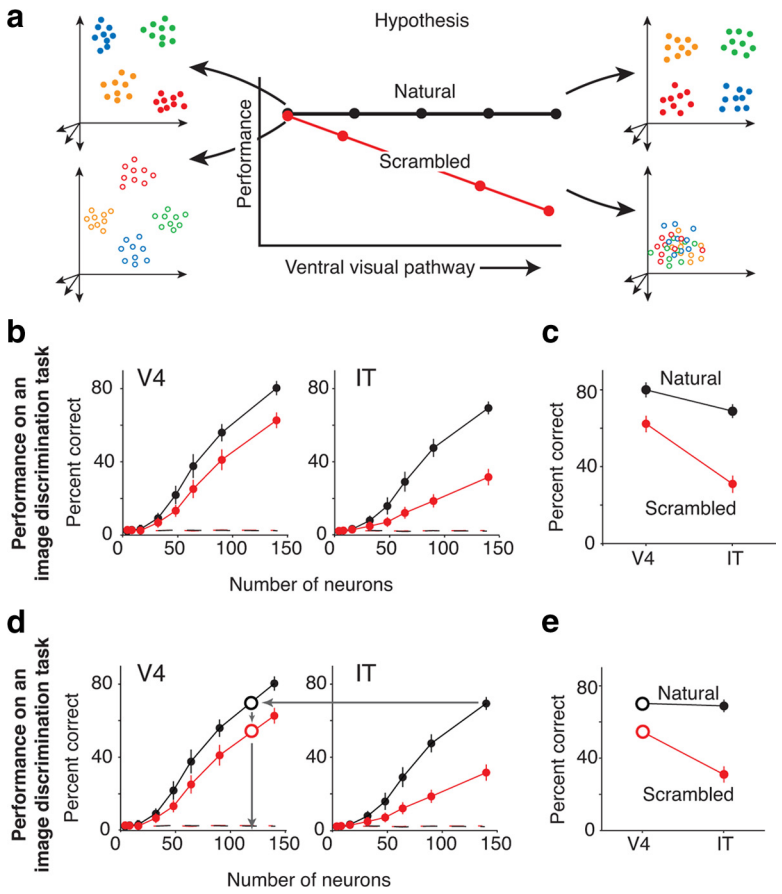

Figure 5. Testing conjunction sensitivity. ***a***, Logic behind the experiment designed to measure conjunction sensitivity. Top left, Response clouds (see Fig. 3) corresponding to the population response to four natural images for an idealized population that encodes local structure within the images. Bottom left, Response clouds for the same population in response to four scrambled versions of the same natural images. In this scenario, scrambling the images activates the population differently, resulting in a repositioning of the response clouds, but the clouds remain a similar distance apart. Top right, Response clouds for an idealized population that encodes specific conjunctions of local structure. Bottom right, Response clouds for the same population in response to scrambled images. In this scenario, destroying the natural feature conjunctions results in the response clouds collapsing toward the origin. ***b***, Performance as a function of the number of neurons for the V4 and IT populations on the discrimination task for the natural (black) and scrambled (red) image sets. Both sets contained 50 images. SE bars indicate the variability (determined by bootstrap) that can be attributed to the specific subset of trials determined for training and testing and the specific subset of neurons chosen. Also shown is chance performance, calculated by scrambling the image labels (dashed lines, ~2%; see Materials and Methods). ***c***, Performance of the V4 and IT populations for $n = 140$ neurons. ***d***, In contrast to equating the number of neurons in each population, V4 and IT can be equated via performance on the natural image set; this amounts to limiting the V4 population to 121 neurons compared with 140 neurons in IT. ***e***, Performance of the V4 and IT populations for $n = 121$ and $n = 140$ V4 and IT neurons, respectively.

Table 1. Scrambling sensitivity

	V4	IT	IT gain
SVM	0.26	0.54	+108%
SVM (matched natural performance)	0.22	0.54	+145%
SVM (subject 1)	0.27	0.46	+70%
SVM (subject 2)	0.29	0.55	+90%
Correlation-based classifier	0.20	0.48	+140%
Normalized Euclidian distance	0.05	0.14	+180%
SVM (25 ms)	0.18	0.60	+230%
SVM (50 ms)	0.36	0.74	+105%
SVM (100 ms)	0.26	0.65	+150%

Scrambling sensitivity measured as the ratio of the performance on the scrambled and natural image discrimination tasks (Fig. 5), subtracted from 1. Included are scrambling sensitivity estimates based on the performance of the linear classifier analysis including the neurons recorded from both subjects and when spikes were counted in a 218 ms window (the duration of each stimulus) and for matched numbers of neurons (Fig. 5*c*), results for numbers of neurons adjusted for matched performance for natural images (Fig. 5*e*), linear classifier performance for subjects 1 and 2 individually, the correlation-based classifier, normalized Euclidean distance metric, and the linear classifier analysis when spikes were counted in 25, 50, and 100 ms windows.

ically, this result demonstrates that we have sampled the visual representation of V4 and IT equally well (by the operational criterion of encoding natural images) and provides a baseline for comparison with the scrambled versions of those images. To determine the sensitivity of the V4 and IT populations to image scrambling, we measured the ability of each population to discriminate among the scrambled versions of the natural images. When V4 and IT are equated in terms of numbers of neurons ($n = 140$), we found moderate reductions in discriminability performance for scrambled images compared with natural images in V4 (scramble performance, 60 ± 4% vs natural performance, 81 ± 4%) and marked decrements in IT (scrambled, 33 ± 4% vs natural, 72 ± 4%) (Fig. 5*b*,*c*). When V4 and IT were equated in terms of their performance on the natural image set (which amounted to limiting the V4 population to 121 neurons compared with 140 neurons in IT), IT still showed a larger deficit in encoding scrambled images than V4 (Fig. 5*d*,*e*).

We summarize these results by computing a measure of scrambling sensitivity as the ratio of the performance on the scrambled and natural image discrimination tasks subtracted from 1. This index takes on a value of 0 for a population that discriminates equally well among natural images and among scrambled images (and is thus insensitive to the particular conjunctions with which the features are presented) and a value of 1 for a population that has absolutely no discriminability for the scrambled image set. With this measure of scrambling sensitivity, we find an increase of 108% in IT over V4 (scrambling sensitivity indices of 0.26 vs 0.54 in V4 and IT, respectively) (Table 1). We observed similar, large increases in the neuronal data from each of our two subjects, confirming that these results are not particular to one individual and were not particular to the fixation behavioral task or the object detection behavioral task (Table 1).

To consider the possibility that these results were peculiar to our SVM linear readout discriminability test, we analyzed the same dataset using two alternative techniques to measure population discriminability. The first method was a simplified, correlation-based classifier that weighted the neurons based on their actual firing rates rather than computing the weights via an optimization process (see Materials and Methods). The second method assessed the average pairwise Euclidean distance between the response clouds (see Materials and Methods). Both methods confirmed that the IT population has substantially higher scrambling sensitivity than V4 (Table 1). Similar results were also found when spikes were counted in shorter windows (i.e., 25, 50, and 100 ms) (Table 1).

These results are consistent with the previously proposed hypothesis that neurons at lower levels of visual processing encode

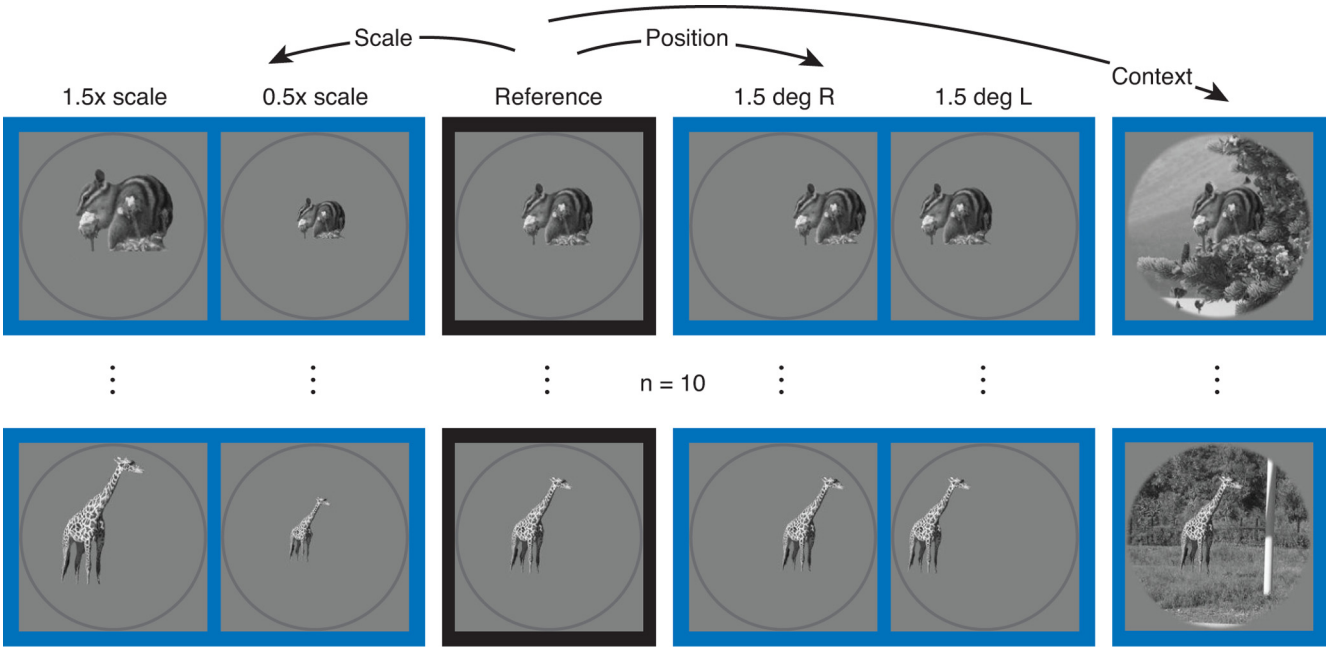

Figure 6. Images used to compare tolerance in V4 and IT. Ten objects were presented under six different transformed conditions. The reference objects (black) were always presented near the center of the 5° aperture. The transformed conditions (blue) included rescaling to 1.5× and 0.5× at the center position, presentation at 1× scale but shifted 1.5° to the right (R) and left (L), and presentation at the reference position and scale but in the context of a natural background.

more local structure whereas neurons at higher stages of the visual system become more sensitive to specific conjunctions of those local features, at least at the sizes of images we presented and the scales at which we scrambled the images. Furthermore, these results suggest that neurons in IT do not merely encode any arbitrary configuration of local structure, rather the IT population is tuned for (i.e., best encodes) the particular configurations found in natural images.

Comparison of tolerance (invariance) in V4 and IT

Next, we used similar population readout approaches to compare the tolerance with object-identity-preserving image transformations between V4 and IT. We used the term "tolerance" instead of "invariance" because it better reflects the fact that object recognition behavior and IT neuronal object recognition performance (Hung et al., 2005; Li et al., 2009) are somewhat insensitive, but not absolutely so, to identity-preserving image transformations. Here, we focus specifically on changes in the position, scale, and background of an object (Hung et al., 2005). Ten objects were presented under six identity-preserving transformations: in addition to a reference condition, which corresponded to an object approximately centered in the 5° aperture on a gray background, the object was shifted left and right, presented at a smaller and larger size, and presented in a natural context (see Materials and Methods) (Fig. 6). We began by training each population to discriminate between objects at the fixed reference condition (Fig. 7*a*, black). Similar to that described above for comparison of selectivity, we regarded discriminability to the reference images as the baseline estimate of the encoding performance of each population. We then asked how well this representation could generalize to the same images presented at the other positions, scales, and at the other background condition (Fig. 7*a*, blue). If the clouds of response vectors corresponding to different transformations of an image remain segregated according to identity, this will translate into good generalization performance (Fig. 7*a*, right). Conversely, if clouds corresponding to the transformed images intermingle with those corresponding to different objects or become located in a completely new location, this will result in poor generalization performance for these identity-preserving transformations (Fig. 7*a*, left). Notably, in this scenario, poor generalization results not from a lack of an ability of the population to encode the individual images themselves but because the identity information about one object is "tangled" with identity information about other objects (DiCarlo and Cox, 2007).

Before probing the ability of the population to generalize across identity-preserving transformations, it was important to ensure that each image corresponding to each transformation was itself well represented by the population. For example, if we had failed to record from V4 receptive fields that tiled the left side of the image, we may have failed to encode the "shift-left" condition altogether and poor generalization would result trivially. Thus, we first tested the ability of the population to discriminate between the 10 objects under each transformed condition separately, and, although we found some variability, in all cases discriminability was high. Notably, consistent with our results on encoding natural images (see above and Fig. 5), we found similar encoding performance for these "natural" objects in V4 and IT (Fig. 7*b*). Specifically, drawing on the same number of neurons in each area ($n = 140$), we found that, on average, V4 and IT encoded the object-related information in each of the transformed conditions with similar, high performance (mean magnitude performance in V4 and IT was 79 and 83% across all six conditions).

Having established that all the individual images were encoded by the V4 and IT populations we recorded, we then asked how well the format of the information in each population was suited for an invariant object recognition task. Specifically, we assessed generalization capacity in V4 and IT using linear classifier methods (Fig. 7*a*). When asked to generalize across small changes in position (from the reference to shifts right or left of 1.5°), the V4 population performance was above chance but decreased markedly relative to the reference (V4 generalization

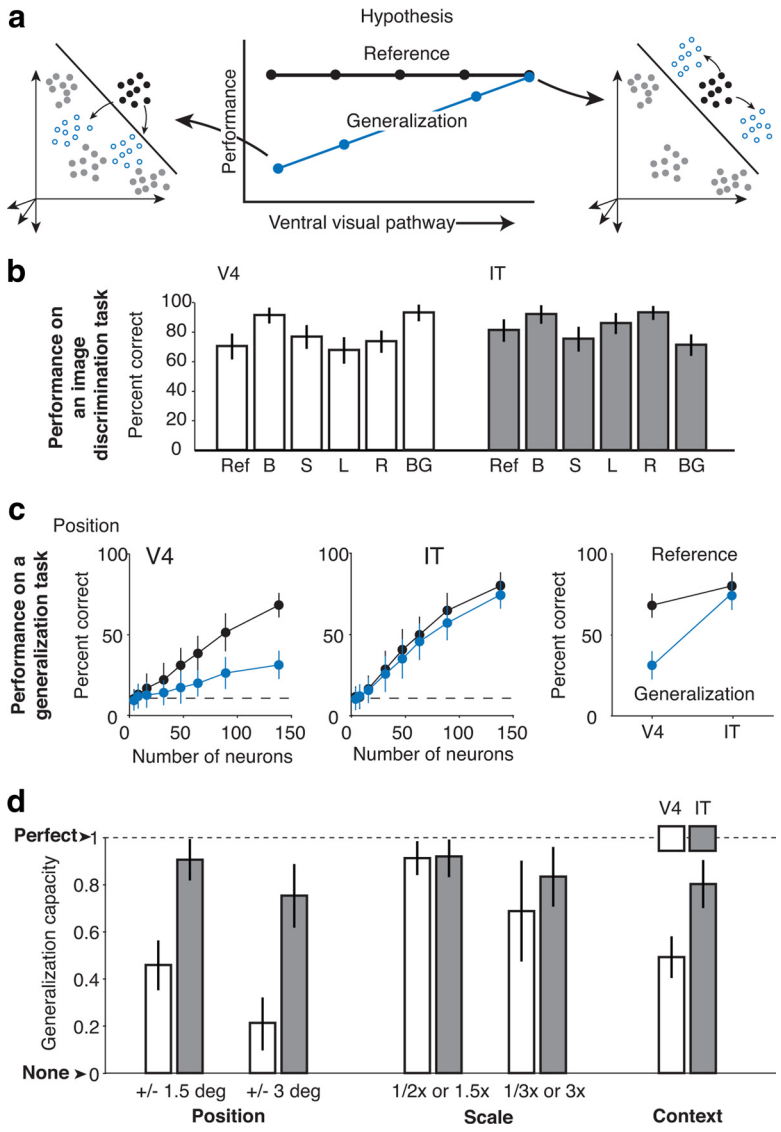

Figure 7. Comparing tolerance in V4 and IT. ***a***, Logic behind the experiment. The analysis begins by training the linear readout to identify the reference objects and then determining how well this representation generalizes across different positions, scales, and context. Middle, We are testing the hypothesis that the ability to generalize across identity-preserving transformations increases along the pathway. Left, More specifically, we expect that neural populations at earlier stages of visual processing will not be capable of generalization because the response clouds for the images presented at different positions, scales, and context will intermingle with the response clouds for other objects, resulting in reduced discriminability. Right, Conversely, we expect that neural populations at later stages of processing will be capable of generalization because the response clouds for the images presented at different positions, scales, and context will remain on the "correct" side of the linear decision boundary. ***b***, To first assess how well the individual images were encoded, performance on the object discrimination task was determined by training and cross-validated testing on different trials of the same images (similar to the black lines in Fig. 5*b*). Plotted is the mean performance on the object discrimination task (chance, 10%). Error bars indicate SEs (determined by bootstrap) that can be attributed to the specific subset of trials determined for training and testing and the specific subset of neurons chosen. Ref, Reference; B, 1.5× scale (Big); S, 0.5× scale (Small); L, 1.5° shift left; R, 1.5° shift right; BG, presentation on a natural background. Performance was high across all transformations in both V4 and IT. ***c***, Generalization across position for the V4 (left) and IT (middle) populations. Black lines indicate mean performance as a function of the number of neurons when training and testing on the reference objects. Blue lines indicate average performance when asked to generalize across small changes in position (from the reference to 1.5° to the left or right). Dashed lines indicate chance performance (~10%), calculated by scrambling the image labels (see Materials and Methods). Error bars indicate SEs (determined by bootstrap) that can be attributed to the specific subset of trials determined for training and testing and the specific subset of neurons chosen. Right, Performance of the V4 and IT populations when nearly all recorded neurons ($n = 140$) from each area are included. ***d***, Generalization capacity for different transformations, calculated as the fractional performance on the generalization task relative to the reference. For example, generalization capacity across small changes in position (the 2 leftmost bars) is calculated as the ratio of the blue and black points in ***c*** (right). Large changes in position correspond to the average generalization across 3° transformations (right to left and left to right): small changes in scale correspond to the average generalization from the reference to the 0.5× and 1.5× images; large changes in scale correspond to the average generalization from 0.5× to 1.5× and vice versa; and changes in context correspond to average generalization from objects on a gray to natural background and vice versa.

performance, 32 ± 9% vs reference performance, 69 ± 8%) (Fig. 7*c*, left). In comparison, the IT population showed much better generalization performance across the same position changes, almost as good as the reference condition (IT generalization performance, 74 ± 9% vs reference, 81 ± 8%) (Fig. 7*c*, middle). To compare across different object-identity-preserving image transformations, we measured generalization capacity as the ratio between performance on each generalization condition and performance on the reference condition (a value of 1.0 indicated perfect generalization capacity) (Fig. 7*d*). We found larger generalization capacity in IT compared with V4 for small (V4, 0.46 ± 0.11; IT, 0.92 ± 0.08) and large (V4, 0.21 ± 0.11; IT, 0.76 ± 0.14) changes in position as well as changes in background context (V4, 0.49 ± 0.09; IT, 0.81 ± 0.10) (Fig. 7*d*), indicating generalization capacity was higher in IT compared with V4 for these transformations. Although small changes in scale resulted in good generalization capacity in both V4 and IT (V4, 0.89 ± 0.09; IT, 0.90 ± 0.09) (Fig. 7*d*), large changes in scale showed a trend toward higher generalization capacity in IT over V4 (V4, 0.69 ± 0.21; IT, 0.83 ± 0.13) (Fig. 7*d*). As a summary statistic, we computed the mean generalization capacity across all five (equally weighted) transformations and found it to be smaller in V4 compared with IT (V4, 0.55; IT, 0.84). These results directly demonstrate that the ability of the IT population to encode object identity is more tolerant of identity-preserving image transformations than the V4 population.

Additional analyses confirmed that these increases in generalization capacity in IT over V4 were found in both monkeys, indicating that they were not peculiar to one individual and were not peculiar to the fixation behavioral task or the object detection behavioral task (Table 2). Similar results were found with the simpler correlation-based classifier readout procedure, indicating that they were not peculiar to the SVM optimization (Table 2). Similar results were also found when spikes were counted in shorter windows (i.e., 50 and 100 ms) (Table 2). To determine whether the variability introduced by the small deviations in eye position across repeated presentations of the same stimulus could account for reduced generalization capacity in V4 compared with IT, we calculated the mean eye position during the presentation of each stimulus and included only those stimulus presentations for which eye position (rel-

ative to the mean position across all trials) fell within a 0.2° diameter window. We calculated the average firing rate response to each stimulus from five such trials, and we disregarded the small proportion of neurons (~5% in V4 and IT) for which five trials failed to meet this criteria for all 60 images. Finally, we reintroduced trial-to-trial variability via Poisson spiking simulations centered on those empirically observed mean rates and recalculated generalization capacity (for $n = 130$ neurons in each population) as described above. Generalization capacity remained higher in IT than V4 (V4, 0.36; IT, 0.55), showing that lower generalization capacity in V4 is unlikely to be caused by the variability introduced by eye movements.

Thus far, we have compared the ability of the V4 and IT populations to generalize across position, scale, and context by training the linear readout on the responses to the reference images and measuring how well these representations generalize to new conditions, and we found that IT performed better at this generalization task than V4. However, this leaves open the possibility that the V4 population representation has the capacity to directly support object identification tasks, even in the face of identity-preserving image variation, but the particular classification training method we have used failed to find it. For example, perhaps the V4 representation can, in principle, allow a hyperplane to separate all images of each object from all other object images but, in practice, this linear separation (hyperplane) is not discovered by only training on the reference image (Fig. 8*a*). That is, although the issues of generalization performance (which we tested in Fig. 7) and linear separability (Li et al., 2009) are related (e.g., good generalization performance with linear classifiers implies linear separability), lack of generalization does not necessarily imply lack of linear separability.

Thus, to directly measure linear separability of the V4 and IT populations, we performed an additional analysis in which we trained on a subset of neuronal population data from all six transformations of each object simultaneously and tested with cross-validated data (i.e., all image conditions were used for training, but the performance was tested on stimulus repetitions that were not used in the training procedure). We found that IT also performed better on this invariant object recognition task than did V4 (Fig. 8*b*, solid lines). Importantly, enhanced performance in IT over V4 is not attributable to an enhanced representation of each image individually because the individual images were represented approximately equally well in V4 and IT (Fig. 7*b*). Furthermore, when we attempted to find hyperplanes that separated arbitrary groups of images (e.g., grouping the reference image of the squirrel with the left-shifted image of a bottle and the enlarged image of a gorilla, etc.) from all other images, we found that IT (and V4) performance dropped nearly to chance (Fig. 8*b*, dashed lines). This shows that, although a relatively small IT population (here $n = 140$) is highly capable of linearly separating image groups produced by object-identity-preserving transformations (Fig. 8*b*, solid lines), it is not capable of supporting arbitrary linear separation of image groups (Fig. 8*b*, dashed lines). In summary, these results directly show that the format of the IT representation is superior to V4 with regard to invariant object identification tasks

Table 2. Mean generalization capacity

	V4	IT	IT gain
SVM	0.55	0.84	+53%
SVM (subject 1)	0.61	0.91	+49%
SVM (subject 2)	0.64	0.88	+38%
Correlation-based classifier	0.58	0.92	+63%
SVM (25 ms)	N/A	0.74	N/A
SVM (50 ms)	0.65	0.75	+15%
SVM (100 ms)	0.61	0.94	+54%
Eye deviation within 0.2°, Poisson variability	0.36	0.55	+53%

Mean generalization capacity across all identity-preserving transformations (Fig. 7*d*). Included are mean generalization capacity estimates based on the performance of: the linear classifier analysis including the neurons recorded from both subjects and when spikes were counted in a 218 ms window (the duration of each stimulus), the linear classifier performance for subjects 1 and 2 individually, the correlation-based classifier, the linear classifier analysis when spikes were counted in 25, 50, and 100 ms windows, and the linear classifier analysis when mean firing rates were computed across trials on which eye position deviated <0.2° and trial-to-trial variability was simulated with a Poisson process (see Results). In 25 ms windows, performance of the V4 population on some of the images (Fig. 7*b*) was not significantly different from chance and thus generalization could not be assessed.

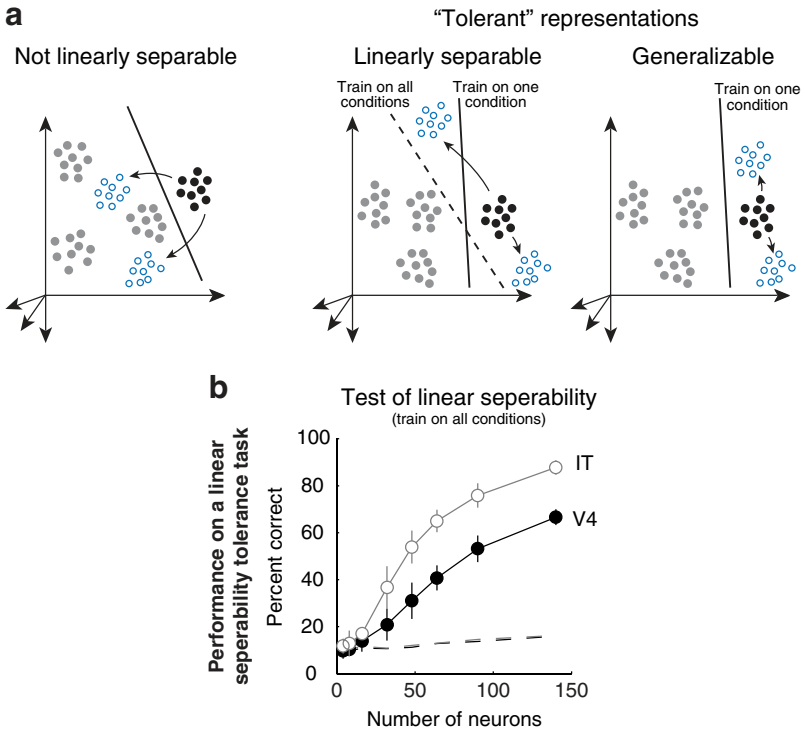

Figure 8. A second tolerance test: linear separability. ***a***, Hypothetical representations that perform poorly and well at tests of tolerance. Left, A population representation that will fail to support tolerant object recognition because of a lack of linear separability. Middle, A population representation that can, in principle, support tolerant object recognition, but one for which the generalization test presented in Figure 7 may fail to find it. The linear boundary located by training on the reference condition alone (solid line) fails to separate the response vectors corresponding to different transformations of one object from the response vectors corresponding to the other objects. In this case, a more appropriate linear boundary (dashed line) can be located by training on all transformations simultaneously. Right, A population representation that performs well at tests of tolerance when training on the reference condition alone. ***b***, Performance on an object identification task for the six transformations of each of 10 objects when the linear boundaries were trained on response data from all six transformations simultaneously and tested with cross-validation data in IT (white) and V4 (black). Dashed lines indicate performance when different objects at different transformations are randomly assigned to one another (e.g., object 1 at the reference position and scale paired with object 3 shifted left 1.5° and object 6 at 0.5× scale, etc.).

(in terms of both linear separability and generalization), and these results are not simply explained by a lack of information in V4 (Fig. 7*b*).

What single-unit properties underlie the increase in population selectivity from V4 to IT?

Our experiments were designed to probe selectivity and tolerance at the population level as opposed to single-neuron measurements; single-neuron experiments typically involve tailoring the stimulus identity and location to match the receptive field of each neuron, whereas our experiments aimed to directly assess real-world tasks by testing the exact same stimulus conditions for each neuron. However, the performance of our population must result from the single neurons contained therein, and, thus for a complimentary look at our data, we computed single-neuron measures. We began by assessing how well individual neurons could discriminate between different natural images and between different scrambled images (Fig. 4). To do this, we used an ROC analysis (see Materials and Methods) to measure the discriminability between all pairs of natural images in our set (1225 pairs) and discriminability between all pairs of scrambled images in our set (1225 pairs). Mean pairwise single-neuron ROC for natural images was not statistically different between V4 and IT (mean ROC: V4, 0.648; IT, 0.639; $p = 0.10$). However, mean pairwise single-neuron ROC was slightly lower for scrambled images than for natural images in V4 (mean ROC: natural, 0.648; scramble, 0.637; $p = 0.04$) and distinctly lower in IT (mean ROC: natural, 0.639; scramble, 0.617; $p < 0.0001$). Thus, our measures of scrambling sensitivity assessed at the single-neuron level are in qualitative agreement with our measurements of discriminability using population-based analyses.

To determine the degree to which the distribution of single-neuron ROC values across the V4 and IT populations contributed to the population-based results for scrambling sensitivity (Fig. 5), we organized our neurons in rank-order according to their ROC values and computed population performance for sliding window subpopulations of 48 neurons with neighboring ROC values. As expected, performance increased with increasing ROC in V4 and IT (Fig. 9*a*). Moreover, Figure 9*a* shows that, when subpopulations of V4 and IT neurons with the same mean single-neuron ROC values are compared, they give rise to the same population performance for natural and scrambled image discriminability (Fig. 9*a*), suggesting that the population-based effects we observed can be directly explained by single-neuron discriminability.

We were also interested in knowing how the distribution of receptive field sizes within each area, and the larger receptive field sizes in IT compared with V4, impacted population performance for natural and scrambled images. To address this question, we began by computing an estimate of receptive field size using the data collected for the invariance experiment in which objects were presented at the center of gaze, 1.5° to the left of center, and 1.5° to the right (Fig. 6). For each neuron, we screened for objects that produced a response significantly differently from baseline, and, for all such objects (if any), we computed an RF profile normalized to the preferred position of each neuron. We quantified individual neuronal position insensitivity as the average fractional response at the two non-optimal positions, and, in agreement with previous studies (Kobatake and Tanaka, 1994), this measure of RF size was larger in IT than V4 (V4 mean, 0.49; IT mean, 0.69; two-tailed *t* test, $p < 0.0001$) (see Fig. 12*a*). Similar to the ROC analysis described above, we organized our neurons in rank order according to their position insensitivity values and

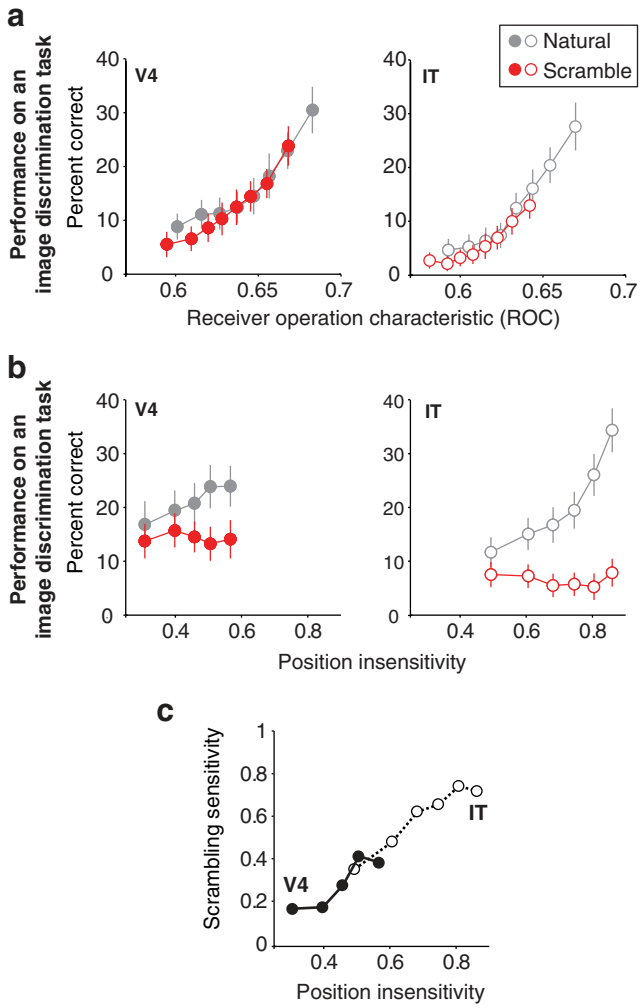

Figure 9. The relationship between single-neuron measures and population discriminability for natural and scrambled images. ***a***, Single-neuron ROC computed for natural (gray) and scrambled (red) images as the average pairwise ROC for 50 natural and 50 scrambled images (Fig. 4). Neurons were ranked separately for their average natural and scrambled image ROC, and population performance for natural or scrambled image discrimination was assessed for subpopulations of 48 neurons with neighboring ROC values. The *x*-axis shows the geometric mean ROC of each subpopulation in V4 (left) and IT (right). The *y*-axis shows performance on the discrimination task for the natural (gray) and scrambled (red) image sets (Fig. 5). ***b***, Single-neuron RF size measured as the insensitivity of the responses to the objects across changes in position (see Fig. 12*a*). Neurons were ranked by position insensitivity, and population performance for natural and scrambled image discrimination was assessed for subpopulations of 48 neurons with neighboring position insensitivity values. The *x*-axis shows the geometric mean position insensitivity of each subpopulation. The *y*-axis shows performance on the discrimination task for the natural (gray) and scrambled (red) image sets (Fig. 5). ***c***, Scrambling sensitivity, calculated as the ratio of scrambled and natural image performance (taken from ***b***), subtracted from one and plotted for subpopulations of 48 neurons with neighboring position insensitivity values.

computed population performance for sliding window subpopulations of 48 neurons with neighboring position insensitivity. For natural images, we found a trend toward increasing performance as RF sizes increased in V4 and a dramatic increase in performance as RF sizes increased in IT (Fig. 9*b*, gray). We found no such trend for the scrambled images in V4 or IT (Fig. 9*b*, red). We also calculated scrambling sensitivity (as described above) as a function of RF size (Fig. 9*c*) and found an approximately monotonically increasing relationship between these two parameters. Moreover, neurons in different visual areas but with similarly sized receptive fields (the largest V4 neurons and the smallest IT

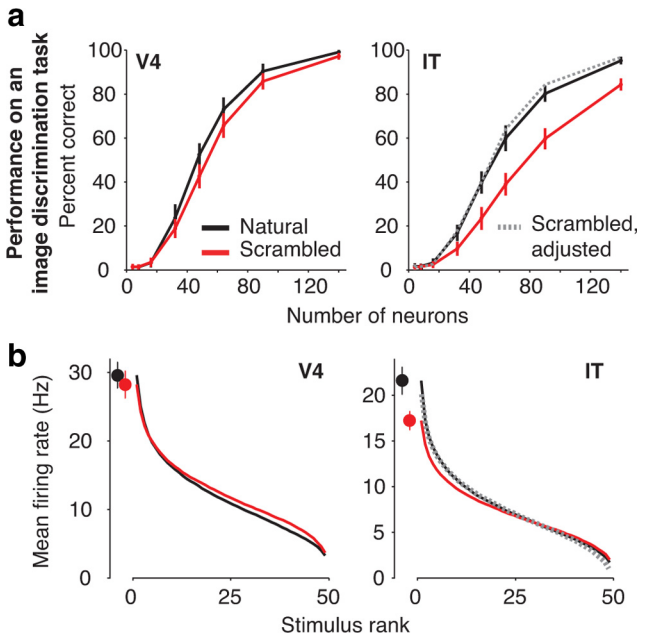

Figure 10. SNR differences can account for decreased scrambled image discriminability in IT. ***a***, Natural (black) and scrambled (red) image discriminability when the mean firing rate of each neuron to each image is preserved but trial-to-trial variability is replaced with a Poisson process. Also shown is discriminability after adjusting the mean dynamic range of the IT population response to scrambled images to match the mean dynamic range of the population response to natural images (gray dashed; see ***b***). ***b***, Mean dynamic range for natural (black) and scrambled (red) images, computed by averaging over the responses of all neurons after organizing the responses of each neuron in rank order. Points on the left of each plot show mean and SE of firing rate to the most effective natural (black) and scrambled (red) stimulus, averaged across all neurons (to indicate error for firing rates of stimulus rank 1). Gray dashed lines indicate the mean dynamic range after the responses of each IT neuron to scrambled images are adjusted with a multiplicative factor and an offset to match the IT responses to natural images (see Results).

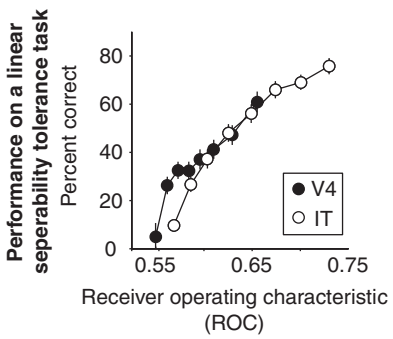

Figure 11. The relationship between single-neuron ROC and population tolerance. Single-neuron ROC computed as the ROC for discriminations between all six transformations of the best object of a neuron (defined by the highest firing rate after averaging across all transformations) and all six transformations of the nine other objects. Neurons were ranked by their ROC values, and population performance for the linear separability test of tolerance (see Fig. 8*b*) was assessed for subpopulations of 48 neurons with neighboring ROC values in V4 (black) and IT (white). The *x*-axis shows the geometric mean ROC of each subpopulation. The *y*-axis shows population performance for the linear separability test of tolerance (see Fig. 8*b*).

neurons) had similar scrambling sensitivities. An interpretation of these results, including the role that neurons with larger RFs might play in encoding natural images is included in Discussion.

For a closer look at the single-neuron response properties that determine population performance for natural and scrambled images, we were interested in knowing whether differences in the signal-to-noise ratio (SNR) of individual neurons for natural versus scrambled images in IT could account for the population behavior or whether something more subtle was happening. We began by maintaining the mean firing rates of each neuron in our population to each image, but we replaced the trial-to-trial variability of each neuron with a Poisson spike generation process. Overall, performance increased in both V4 and IT relative to the raw data, and, although performance for natural and scrambled images was similar in V4, performance for scrambled images remained lower than performance for natural images in IT (Fig. 10*a*). This suggests that the differences in performance of the IT population for natural compared with scrambled images are not simply attributable to differences in trial-by-trial variability for the two image classes. Next we investigated the average dynamic response range of V4 and IT neurons by plotting the rank-order firing rate plots for the 50 natural and 50 scrambled images, averaged across each population (Fig. 10*b*). Consistent with the ROC results presented in Figure 9*a*, the rank-order plots were similar for natural and scrambled images in V4 (Fig. 10*b*, left), whereas natural images produced a slightly larger average dynamic response range than scrambled images in IT (Fig. 10*b*, right). To determine whether the differences in the average IT rank-order curves for natural compared with scrambled images was sufficient to account for the increased population performance, we aligned the dynamic range of the two curves by applying the same multiplicative factor and offset to the rank-order response curve of the responses of each IT neuron to scrambled images ($R_{new} = 1.27 \times R_{orig} - 1.65$) and, as before, simulated variability with a Poisson spike generation process. Performance of this "adjusted" IT population to scrambled images matched the population performance to natural images (Fig. 10*a*, gray dashed), suggesting that the differences in SNR between natural and scrambled images can account for the differences we observed in population performance for these two image classes.

What single-unit properties underlie the increase in population tolerance from V4 to IT?

Similar to our inquiry of the single-neuron correlates of population scrambling sensitivity, we were interested in the single-neuron correlates of population tolerance. We began by computing the single-neuron ROC for discrimination between all six transformations of the best object of a neuron and all transformations of the nine other objects; we found that single-neuron ROC was higher in IT than V4 (mean ROC: V4, 0.610; IT, 0.659; $p < 0.0001$). Next, we organized our neurons in rank order according to their ROC values and computed population performance on an invariant object recognition task (the task described for Fig. 8*b*) for sliding window subpopulations of 48 neurons with neighboring ROC values. The curves describing population performance for tests of linear separability across the different transformations of each object were nearly aligned in V4 and IT for populations with similar ROC values (Fig. 11), suggesting that single-neuron performance on the tolerance task assessed by ROC is a good predictor of population performance on the same task assessed by linear classifiers.

For a closer look at how the sizes of individual receptive fields impacted population tolerance, we took a slightly different approach from that described for scrambling sensitivity. Specifically, we wanted to understand the degree to which the higher performance on the invariant object recognition task observed in IT over V4 (Fig. 8*b*) was correlated with larger IT RFs. To address this question, we began by estimating the position sensitivity of V4 and IT neurons as described above (i.e., an estimate of RF size)

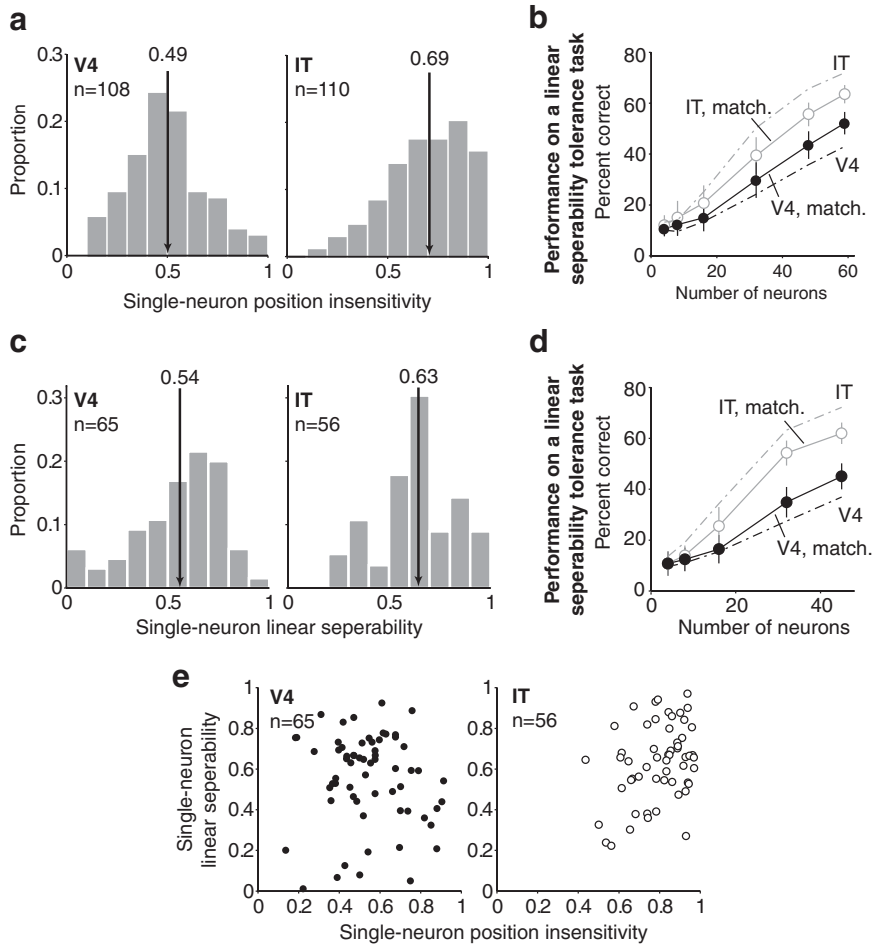

Figure 12. Single-neuron correlates of population tolerance. ***a***, Position insensitivity of the V4 and IT population based on the responses to the objects presented at the center of gaze, 1.5° to the left of center, and 1.5° to the right (see Fig. 6). For each neuron, a receptive field profile was computed as the average response at each position to all objects that produced a response significantly differently from baseline at one or more positions (*t* test, $p < 0.05$; 108 of 140 neurons in V4 and 110 of 143 neurons in IT). After normalizing the receptive field profile to 1 at the preferred position, we quantified position sensitivity as the average fractional response at the two non-optimal positions. Arrows indicate means. ***b***, Performance on the same object identification task presented in Figure 8*b* but for an IT and V4 population that are matched for position insensitivity. Each subpopulation was chosen by randomly sampling the maximal number of entries in each histogram bin in ***a*** that overlapped. Solid lines indicate populations matched in this way in IT (white) and V4 (black). Dashed lines indicate populations that passed the significance test for at least one object but when not equated for position sensitivity (all the entries in the histograms of ***a***). ***c***, Single-neuron linear separability index measured as the correlation between the actual responses of the neuron and the predicted responses assuming independence between the responses to the 10 objects and each of six transformed conditions (see Fig. 6 and Materials and Methods). For this analysis, only neurons that responded significantly differently from baseline to at least one object under at least two transformed conditions were included (V4, $n = 65$ of 140; IT, $n = 56$ of 143). Arrows indicate means (V4, 0.54; IT, 0.63). ***d***, Similar to ***b***, performance on the same object identification task presented in Figure 8*b* but for an IT and V4 population that are matched for their linear separability index ($n = 45$; mean V4, 0.586; IT, 0.588). Solid lines indicate populations matched in this way in IT (white) and V4 (black). Dashed lines indicate populations that passed the significance test for at least one object but when not equated for single-neuron linear separability (all the entries in the histograms of ***c***). ***e***, Plots of position insensitivity versus linear separability index in V4 (left) and IT (right).

(Fig. 12*a*). To determine whether the differences in RF size between V4 and IT could account for the differences we observed in population performance for linear separability, we subselected the largest possible V4 and IT subpopulations that were matched for average RF position insensitivity ($n = 59$). The effect of equating V4 and IT for position sensitivity resulted in population performance in the invariant object recognition task that was more similar in V4 and IT compared with the differences observed with the entire population, but performance remained significantly higher in IT (Fig. 12*b*). Thus, the differences in single-neuron RF size alone cannot account for the differences in linear separability observed between these two populations.

What single-unit property then can account for higher linear separability of object identity in the IT population? In contrast to absolute position sensitivity, a second concept of "receptive field" relies on the notion that a neuron that contributes to a highly tolerant object representation at the population level will maintain its relative selectivity (rank-order) for different objects across changes in position, scale, and context, although the absolute responses of the neuron might rescale with each transformation (Tovee et al., 1994; Ito et al., 1995; Logothetis and Sheinberg, 1996; Op De Beeck and Vogels, 2000; DiCarlo and Maunsell, 2003; DiCarlo and Cox, 2007). Extensive simulations have established a link between rank-order preservation of selectivity and population performance under mild assumptions about the distribution of tuning in the population (Li et al., 2009). To measure rank-order preservation of selectivity in individual neurons, we began by screening for neurons that responded significantly differently from baseline to at least one of the objects across two (of six) transformations (Fig. 6) (see Materials and Methods). For these neurons, we quantified how well rank-order selectivity was preserved across changes in position, scale, and context using a single-neuron linear separability index (Mazer et al., 2002; Brincat and Connor 2004; Janssen et al., 2008; Li et al., 2009) that measures how well the response of the neuron can be explained as resulting from independent tuning for object identity and changes in position, scale, and context. This correlation-based index compares the similarity of the actual response surface with the best-fitting response surface assuming independent tuning and takes on a value of 1 for perfect independence (see Materials and Methods). Although the metric varies widely from neuron to neuron, we found that, on average, IT had significantly higher single-neuron linear separability index values when compared with V4 (Fig. 12*c*) (mean V4, 0.54; mean IT, 0.63; two-tailed *t* test, $p = 0.0263$). To determine whether this difference in single-neuron linear separability between the two areas could account for the difference we observed in population performance on our invariant object recognition task (Fig. 8), we subselected from our populations the largest possible V4 and IT subpopulations that were matched for this measure ($n = 45$; mean V4, 0.586; mean IT, 0.588). Similar to equating RF position sensitivity, the effect of equating V4 and IT for single-neuron linear separability resulted in population performance on an invariant object recognition task that was more similar between V4 and IT, but performance remained significantly

higher in IT (Fig. 12*d*). Thus, differences in the maintenance of relative selectivity across changes in position, scale, and context between V4 and IT also cannot completely account for the differences in population performance between these two visual areas.

Although there is no theoretical reason to suggest that these two single-neuron measures of tolerance (position sensitivity and rank-order preservation of selectivity) should be correlated with one another, we wondered whether the neurons with the largest receptive fields also tended to be those that best preserved their rank-order selectivity across identity-preserving transformations. We found that RF position sensitivity was not significantly correlated with our single-neuron linear separability index in V4 (Fig. 12*e*, left) ($r = -0.16$; $p = 0.52$) and was weakly but significantly correlated with our single-neuron linear separability index in IT (Fig. 12*e*, right) ($r = 0.34$; $p = 0.01$). Thus, although both the average RF position sensitivity and rank-order selectivity increase from V4 to IT, these two single-unit measures are not empirically bound together. This suggests the hypothesis that increases in both RF position sensitivity and the preservation of rank-order selectivity may be required to match V4 and IT performance for invariant object recognition tasks.

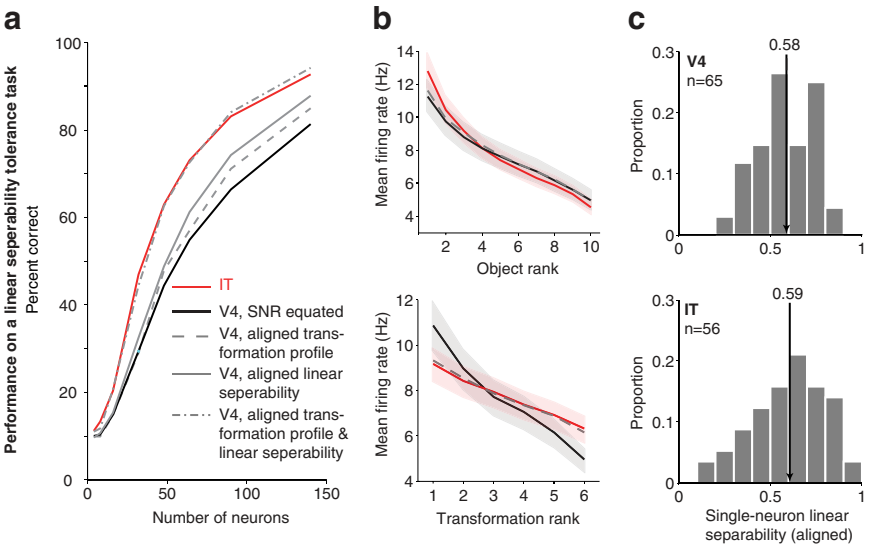

Figure 13. Receptive field differences can account for higher population tolerance in IT. ***a***, Performance of the V4 (black) and IT (red) populations on the invariant object recognition task described for Figure 8*b* after equating V4 and IT for SNR (responses of each V4 cell were rescaled by 0.78, and trial-to-trial variability of both V4 and IT neurons was simulated by a Poisson process). Additional transformations to the V4 population include the following: aligning the V4 transformation profile to match the average IT transformation profile (gray dashed; see Results; ***b***); aligning the average V4 single-neuron linear separability to match the average in IT (gray solid; see Results; ***c***); aligning both the V4 transformation profile to match the average IT transformation profile and aligning the average V4 single-neuron linear separability to match the average in IT (gray dot-dashed). ***b***, Top, Average across the population of the rank-order response to all objects computed after averaging across all six transformations. Bottom, Average across the population of the rank-order response to all six transformations computed after averaging across all 10 objects. V4, Black; IT, red; gray dashed, V4 after aligning the transformation rank profile to match IT. Colored regions indicate mean ± 1 SE. ***c***, Single-neuron linear separability histograms computed after manipulating the internal structure of the V4 neurons' 10 object × 6 transformation response surface to match the average single-neuron linear separability in IT (see Results). Arrows indicate means.

To more closely examine this hypothesis, we sought to determine a set of manipulations that could transform the V4 population into one that was matched to IT for both average single-neuron metrics and population performance on an invariant object recognition task. Similar to the procedure described in Figure 10, we began by equating the SNR in the two populations. V4 neurons had higher average firing rates than IT (data not shown); hence, we aligned the average V4 and IT firing rate responses to the 60 stimuli (10 objects each presented under six transformations) by applying the same multiplicative factor (0.78) to the responses of each V4 neuron and simulated trial-to-trial variability with a Poisson spike generation process. IT performance remained higher than V4 (Fig. 13*a*, red versus black), suggesting that differences in SNR alone cannot account for increased IT performance. Note that, for all analysis described below, we re-equate SNR in a similar manner.

To determine whether single-neuron differences in firing rate sensitivity to identity-preserving transformations [which include position sensitivity ("RF"), described above, as well as scale and background sensitivity] could account for the differences in population performance between V4 and IT, we examined the mean firing rate to each object (collapsed across all six identity-preserving transformations) and the mean firing rate to each transformation (collapsed across all 10 objects) after organizing the responses along each axis in rank order for each neuron. We found that rank-order plots averaged over all neurons in V4 versus IT were similar for the 10 objects (Fig. 13*b*, top) but were substantially different across six transformations (Fig. 13*b*, bottom). Specifically, the mean V4 "transformation profile" fell off more steeply away from the preferred transformation, whereas the mean IT transformation profile was more flat. We equated the average transformation profile in V4 to IT by applying the same multiplicative adjustment for each V4 neuron to each rank-ordered transformation (weights of 0.75, 0.89, 0.95, 0.99, 1.06, and 1.4); this manipulation resulted in a small increase in V4 population performance but V4 performance remained lower than in IT (Fig. 13*a*, gray dashed). Consistent with Figure 12*b*, this suggests that equating the V4 transformation profile alone is not sufficient to achieve IT population performance.

To determine whether higher single-neuron linear separability (Fig. 12*c*) could account for the increased tolerance in IT, we manipulated the average linear separability of V4 neurons to match the average in IT. Specifically, for the subpopulation of V4 neurons that passed the screening test described for Figure 12*c*, we manipulated the internal structure of the response surface of each V4 neuron by taking a weighted sum of the independent and non-independent components determined by singular value decomposition (see Materials and Methods). The same weights were applied to each neuron; the weights were designed to approximately match the mean single-neuron linear separability index in V4 and IT (Fig. 13*c*). Similar to the results for the transformation profile adjustment, adjusting the average linear separability index of the V4 population to match the average IT linear separability index alone was not sufficient to match IT performance (Fig. 13*a*, solid gray). Finally, when we simultaneously matched both the V4 transformation profile (Fig. 13*b*) and the V4 single-neuron linear separability index (Fig. 13*c*) to match the values observed in IT, this resulted in an adjusted V4 population whose performance was matched to IT (Fig. 13*a*, gray dot-dash). These results suggest that both the decrease in single-unit response rate sensitivity to identity-preserving image transformations as well as the increase in single-unit rank-order selectivity preservation

across those image transformations are required (empirically speaking) to account for the ability of the IT population to outperform the V4 population in invariant object recognition tasks. Restated for the specific case of recognizing objects across changes in retinal position, this corresponds to both a broadening of the spatial RF as well as an increase in the rank-order object preference at each location within the spatial RF.

Discussion

Although the hierarchy of visual areas that combine to form the ventral visual stream has been identified through anatomical connections and latency estimates, little is known about how visual information is reformatted as signals pass from one visual area in the stream to the next. To better understand this reformatting, we focused on measuring any change in the two key properties required for object recognition: selectivity and tolerance. Here we focused on a form of selectivity that describes the complexity of image features that activate a neuron, conjunction sensitivity, measured as the sensitivity to natural image scrambling. We found a substantial increase in scrambling sensitivity from V4 to IT, which suggests an increase in sensitivity for conjunctions of local image features, particularly for conjunctions found in natural images. We also found a substantial increase in the tolerance to changes in the position, scale, and context of those feature conjunctions as signals travel from V4 to IT. These increases did not depend on the animal's task or the method of reading out the population code.

Several lines of evidence have already suggested that conjunction sensitivity increases across the ventral visual stream (Desimone et al., 1984; Gallant et al., 1993; Kobatake and Tanaka, 1994; Pasupathy and Connor, 1999; Vogels, 1999; Brincat and Connor, 2004; Anzai et al., 2007; Hegdé and Van Essen, 2007). Notably, Kobatake et al. (1994) searched for the minimally complex feature arrangements that would robustly drive neurons at different points of the ventral visual stream and found indications of a gradient of complexity in terms of an increase in the relative responses to preferred "complex" and "simple" stimuli. Another elegant series of studies has explored tuning for parametrically defined contour shapes (Pasupathy and Connor, 1999; Brincat and Connor, 2004; Yamane et al., 2008) and found indications of increasingly complex contour encoding properties across different stages of the ventral visual stream (as determined by the strength of the nonlinear terms needed to fit responses in a particular curvature basis). Although such single-neuron approaches have the potential to provide insight into the specific visual features that neurons in each area are encoding, definitive determination of increases in conjunction sensitivity requires probing different visual areas with the same stimulus set, presented at the same size and in the same region of the visual field. By monitoring the combined behavior of neurons as a population, we demonstrate that, as information about an image propagates through the ventral pathway, the representation of that image becomes increasingly selective for specific conjunctions of local features found in natural images. These results are consistent with functional magnetic resonance imaging (fMRI) results in monkey extrastriate visual cortex (Denys et al., 2004) and are likely analogous to findings that regions of human high-level visual cortex produce more robust fMRI signals to intact objects compared with scrambled images (Kanwisher et al., 1996, 1997).

The finding that the neuronal population at a later stage of processing has a decreased ability to encode scrambled images shows that information about these images is first encoded by the visual system and later disregarded. That is, successive ventral stream representations behave as if they are becoming increasingly tuned to naturalistic stimuli or stimuli that overlap the region of shape space previously encountered by the animal (and thus implicitly disregard less natural, "scrambled" stimuli). In these experiments, the monkeys were not asked to discriminate between different scrambled images; given the plasticity known to exist at higher stages of visual processing (Baker et al., 2002; Sigala and Logothetis, 2002; Kourtzi and DiCarlo, 2006; Li and DiCarlo, 2008), it remains an open question as to how well IT populations can encode less natural stimuli such as scrambled images after training.

With regard to changes in tolerance (invariance) along the ventral stream, previous studies have illustrated that single IT neurons tend to maintain their rank-order selectivity for images across changes in position and scale (Schwartz et al., 1983; Tovee et al., 1994; Ito et al., 1995; Logothetis and Pauls, 1995; Op De Beeck and Vogels, 2000) and that relatively small IT populations are capable of supporting position and scale invariant object recognition (Hung et al., 2005). The results of one fMRI study suggest a posterior-to-anterior gradient of size tolerance within monkey IT (Sawamura et al., 2005). Our results extend these previous findings to demonstrate that, in addition to position and scale tolerance, the IT population representation is at least somewhat tolerant to context. In addition, our results are the first to directly show that this increase in tolerance happens gradually across different visual areas along the ventral stream. That is, a portion, but not all, of the position, scale, and context tolerance properties observed in IT likely arise from computations that occur in IT itself because they are weaker in V4. Qualitatively similar increases from V4 to IT have been reported for orientation and shape selectivity invariant to the cue that defines the boundaries (i.e., luminance vs motion) (Sáry et al., 1995; Mysore et al., 2006).

We found three complimentary single-neuron correlates of the increases in population scrambling sensitivity we observed in IT over V4. First, single-neuron ROC values were much lower for scrambled images compared with natural images in IT (in contrast to V4). Moreover, directly comparing population performance for neurons with similar ROC values for natural and scrambled images aligned population performance (Fig. 9*a*), suggesting that (not surprisingly) discriminability at the single-neuron level can account for our population differences. Similarly, we found that the lower IT population discriminability for scrambled images could be accounted for by a smaller dynamic range of response to the scrambled images (Fig. 10). Finally, we found that performance for natural image discrimination (but not scrambled image discrimination) increased with increasing receptive field size in both V4 and IT (Fig. 9*b*). This supports the notion that pooling over local features is important for natural image encoding in both V4 and IT.

Can the increases in conjunction sensitivity that we observed be explained simply by increases in RF size (Desimone and Schein, 1987; Gattass et al., 1988; Kobatake and Tanaka, 1994; Op De Beeck and Vogels, 2000) as signals pass from V4 to IT? The answer to this question depends entirely on what one means by an "increase in the receptive field size." Our results rule out models in which receptive field size increases are implemented from V4 to IT by "magnifying" small RFs (e.g., taking a small Gabor and producing a larger scale copy) or by pooling small RFs with similar preferences because both of these alternatives would encode natural and scrambled images with similar fidelity. Importantly, we found a relationship between scrambling sensitivity and RF size in both visual areas and that V4 and IT neurons with similarly sized RFs had similar sensitivities to image scrambling

(Fig. 9*c*). Thus, our results are consistent with models in which the number of conjunctions implemented by a neuron in the ventral pathway is, on average, proportional to the size of its RF. In other words, IT has a higher conjunction sensitivity than V4 and it has correspondingly larger RFs, but we do not know whether one response property causes the other at a deeper mechanistic level. Nor do we yet know whether this phenomenology is somehow optimal in an algorithmic resource allocation sense.

In our experiments, we fixed the size of the images and the four scales at which the scrambling was applied. Importantly, we used natural images presented at sizes within the regime that one would encounter in the natural world, and our results demonstrate that, within this regime, IT is more sensitive to the specific configurations of features found in natural images. Notably, the smallest scale of scrambling we used (filters sized 0.625°) is a fraction of the size of most V4 RFs at the eccentricities we recorded (Desimone and Schein, 1987; Gattass et al., 1988), and thus visual features were in fact "scrambled" within V4 RFs. The modest decreases we observed in V4 at this scale of scrambling suggests that V4 neurons are at least somewhat sensitive to conjunctions of the simple features encoded by our scrambling procedure (Gallant et al., 1993; Pasupathy and Connor, 1999); future investigations will be required to determine whether scrambling at even smaller scales reveals additional conjunction sensitivity in V4.

We were particularly interested in knowing whether the increases in tolerance that we observed in our population-based analyses could be explained by increases in RF sizes between V4 and IT. Again, answering this question requires a careful consideration of what one means by RF. Although larger RF sizes, measured as the position insensitivity for a preferred object, have been documented previously in V4 and IT, we are not aware of any comparison of the degree to which rank-order selectivity is preserved across identity-preserving transformations (which we refer to here as "single-neuron linear separability"). We found that both RF size and single-neuron linear separability are higher in IT than V4 but that these two single-unit measures of an RF were at best weakly correlated within each population. In other words, not all of the largest RF neurons in the population succeed at maintaining their linear separability across changes in position, scale, and context, whereas some of the smaller receptive field neurons maintain their relative selectivity at the low firing rate fringes of their receptive fields. Moreover, we found two pieces of evidence to suggest that neither the increases in transformation bandwidth alone (which includes RF position sensitivity as well as bandwidth for changes in size and context) nor preservation of rank-order selectivity alone could account for the increased invariant object recognition performance of the IT population. First, when we selected V4 and IT subpopulations that were matched for either of these parameters, IT still produced higher performance than V4 (Fig. 12*b*,*d*). Second, using a combined data and simulation analysis, we found that "transforming" the V4 population to match the values of the IT for either parameter alone was not sufficient to account for higher IT performance, whereas matching both parameters together could account for the higher performance observed in IT over V4. The notion that the maintenance of rank-order selectivity preferences across identity-preserving transformations has long been appreciated (Tovee et al., 1994; Ito et al., 1995; Logothetis and Sheinberg, 1996; Op De Beeck and Vogels, 2000; DiCarlo and Maunsell, 2003; DiCarlo and Cox, 2007), and increases in population performance with increased single-neuron linear separability have been predicted previously via simulation (Li et al., 2009). The empirical requirement that transformation bandwidth increase (e.g., increase in spatial RF size) is somewhat less expected. Previous simulations predict that bandwidth is much less important than single-neuron separability in the limit of populations that contain a sufficient number of neurons such that complete "coverage" of the object and transformation space as been achieved (Li et al., 2009). The analyses we present here compare population performance for V4 and IT populations containing the same number of neurons for numbers that are much smaller than the full population. One reasonable interpretation of these results is that, per neuron, the IT population more efficiently "tiles" the transformation space than does V4 by virtue of broader receptive fields across position, size, and context. In summary, increases in tolerance along the ventral pathway are not simply a consequence of first-order increasing RF phenomenology but instead that increasing RF phenomenology is likely a reflection of more sophisticated computations designed for the discrimination of objects invariant of identity-preserving image transformations (Fukushima, 1980; Riesenhuber and Poggio, 1999; DiCarlo and Cox, 2007; Serre et al., 2007).

References

Anzai A, Peng X, Van Essen DC (2007) Neurons in monkey visual area V2 encode combinations of orientations. Nat Neurosci 10:1313–1321.

Baker CI, Behrmann M, Olson CR (2002) Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. Nat Neurosci 5:1210–1216.

Brincat SL, Connor CE (2004) Underlying principles of visual shape selectivity in posterior inferotemporal cortex. Nat Neurosci 7:880–886.

Britten KH, Shadlen MN, Newsome WT, Movshon JA (1992) The analysis of visual motion: a comparison of neuronal and psychophysical performance. J Neurosci 12:4745–4765.

Denys K, Vanduffel W, Fize D, Nelissen K, Peuskens H, Van Essen D, Orban GA (2004) The processing of visual shape in the cerebral cortex of human and nonhuman primates: a functional magnetic resonance imaging study. J Neurosci 24:2551–2565.

Desimone R, Schein SJ (1987) Visual properties of neurons in area V4 of the macaque: sensitivity to stimulus form. J Neurophysiol 57:835–868.

Desimone R, Albright TD, Gross CG, Bruce C (1984) Stimulus-selective properties of inferior temporal neurons in the macaque. J Neurosci 4:2051–2062.

DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. Trends Cogn Sci 11:333–341.

DiCarlo JJ, Maunsell JH (2000) Form representation in monkey inferotemporal cortex is virtually unaltered by free viewing. Nat Neurosci 3:814–821.

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1:1–47.

Fukushima K (1980) Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern 36:193–202.

Gallant JL, Braun J, Van Essen DC (1993) Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. Science 259:100–103.

Gattass R, Sousa AP, Gross CG (1988) Visuotopic organization and extent of V3 and V4 of the macaque. J Neurosci 8:1831–1845.

Green D, Swets JA (1966) Signal detection theory and psychophysics. Wiley: New York.

Hegdé J, Van Essen DC (2007) A comparative study of shape representation in macaque visual areas v2 and v4. Cereb Cortex 17:1100–1116.

Hubel DH, Wiesel TN (1965) Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. J Neurophysiol 28:229–289.

Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005) Fast readout of object identity from macaque inferior temporal cortex. Science 310:863–866.

Ito M, Tamura H, Fujita I, Tanaka K (1995) Size and position invariance of neuronal responses in monkey inferotemporal cortex. J Neurophysiol 73:218–226.

Janssen P, Srivastava S, Ombelet S, Orban GA (2008) Coding of shape and position in macaque lateral intraparietal area. J Neurosci 28:6679–6690.

Kanwisher N, Chun MM, McDermott J, Ledden PJ (1996) Functional imaging of human visual recognition. Brain Res Cogn Brain Res 5:55–67.

Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 17:4302–4311.

Kobatake E, Tanaka K (1994) Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. J Neurophysiol 71:856–867.

Kourtzi Z, DiCarlo JJ (2006) Learning and neural plasticity in visual object recognition. Curr Opin Neurobiol 16:152–158.

Kreiman G, Hung CP, Kraskov A, Quiroga RQ, Poggio T, DiCarlo JJ (2006) Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. Neuron 49:433–445.

Lennie P, Movshon JA (2005) Coding of color and form in the geniculostriate visual pathway (invited review). J Opt Soc Am A Opt Image Sci Vis 22:2013–2033.

Li N, DiCarlo JJ (2008) Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science 321:1502–1507.

Li N, Cox DD, Zoccolan D, DiCarlo JJ (2009) What response properties do individual neurons need to underlie position and clutter "invariant" object recognition? J Neurophysiol 102:360–376.

Logothetis NK, Sheinberg DL (1996) Visual object recognition. Annu Rev Neurosci 19:577–621.

Logothetis NK, Pauls J (1995) Psychophysical and physiological evidence for viewer-centered object representations in the primate. Cereb Cortex 5:270–288.

Mazer JA, Vinje WE, McDermott J, Schiller PH, Gallant JL (2002) Spatial frequency and orientation tuning dynamics in area V1. Proc Natl Acad Sci U S A 99:1645–1650.

Meyers EM, Freedman DJ, Kreiman G, Miller EK, Poggio T (2008) Dynamic population coding of category information in inferior temporal and prefrontal cortex. J Neurophysiol 100:1407–1419.

Mysore SG, Vogels R, Raiguel SE, Orban GA (2006) Processing of kinetic boundaries in macaque V4. J Neurophysiol 95:1864–1880.

Op De Beeck H, Vogels R (2000) Spatial sensitivity of macaque inferior temporal neurons. J Comp Neurol 426:505–518.

Pasupathy A, Connor CE (1999) Responses to contour features in macaque area V4. J Neurophysiol 82:2490–2502.

Portilla J, Simoncelli EP (2000) A parametric texture model based on joint statistics of complex wavelet coefficients. Int J Comput Vis 40:49–71.

Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2:1019–1025.

Ringach DL (2002) Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. J Neurophysiol 88:455–463.

Rolls ET, Tovee MJ (1995) Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. J Neurophysiol 73:713–726.

Sáry G, Vogels R, Kovács G, Orban GA (1995) Responses of monkey inferior temporal neurons to luminance-, motion-, and texture-defined gratings. J Neurophysiol 73:1341–1354.

Sawamura H, Georgieva S, Vogels R, Vanduffel W, Orban GA (2005) Using functional magnetic resonance imaging to assess adaptation and size invariance of shape processing by humans and monkeys. J Neurosci 25:4294–4306.

Schwartz EL, Desimone R, Albright TD, Gross CG (1983) Shape recognition and inferior temporal neurons. Proc Natl Acad Sci U S A 80:5776–5778.

Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T (2007) Robust object recognition with cortex-like mechanisms. IEEE Trans Pattern Anal Mach Intell 29:411–426.

Sigala N, Logothetis NK (2002) Visual categorization shapes feature selectivity in the primate temporal cortex. Nature 415:318–320.

Tanaka K (1996) Inferotemporal cortex and object vision. Annu Rev Neurosci 19:109–139.

Tovee MJ, Rolls ET, Azzopardi P (1994) Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque. J Neurophysiol 72:1049–1060.

Vogels R (1999) Effect of image scrambling on inferior temporal corticla responses. Neuroreport 10:1811–1816.

Yamane Y, Carlson ET, Bowman KC, Wang Z, Connor CE (2008) A neural code for three-dimensional object shape in macaque inferotemporal cortex. Nat Neurosci 11:1352–1360.

Zoccolan D, Cox DD, DiCarlo JJ (2005) Multiple object response normalization in monkey inferotemporal cortex. J Neurosci 25:8150–8164.

Zoccolan D, Kouh M, Poggio T, DiCarlo JJ (2007) Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. J Neurosci 27:12292–12307.