Hue tuning curves in V4 change with visual context

Ari S. Benjamin[1*], Pavan Ramkumar[2, 3], Hugo Fernandes[2], Matthew Smith[4,5,6], Konrad P. Kording[1,2]

1. Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
2. Department of Physical Medicine and Rehabilitation, Northwestern University and Shirley Ryan Ability Lab, Chicago, IL
3. System1 Biosciences, San Francisco, CA
4. Department of Biomedical Engineering, Carnegie Mellon University, Pittsburgh, PA
5. Carnegie Mellon Neuroscience Institute, Pittsburgh, PA
6. Department of Ophthalmology, University of Pittsburgh, Pittsburgh, PA
7. *Correspondence: aarrii@seas.upenn.edu

Summary

To understand activity in the higher visual cortex, researchers typically investigate how parametric changes in stimuli affect neural activity. These experiments reveal neurons' general response properties only when the effect of a parameter in synthetic stimuli is representative of its effect in other visual contexts. However, in higher visual cortex it is rarely verified how well tuning to parameters of simplified experimental stimuli represents tuning to those parameters in complex or naturalistic stimuli. To evaluate precisely how much tuning curves can change with context, we developed a methodology to estimate tuning from neural responses to natural scenes. For neurons in macaque V4, we then estimated tuning curves for hue from both natural scene responses and responses to artificial stimuli of varying hue. We found that neurons' hue tuning on artificial stimuli was not representative of their hue tuning on natural images, even if the neurons were strongly modulated by hue. These neurons thus respond strongly to interactions between hue and other visual features. We argue that such feature interactions are generally to be expected if the cortex takes an optimal coding strategy. This finding illustrates that tuning curves in higher visual cortex may only be accurate for similar stimuli as shown in the lab, and do not generalize for all neurons to naturalistic and behaviorally relevant stimuli.

Keywords: *Encoding model, receptive field, color perception, neurophysiology, tuning curve, V4, visual cortex, natural stimuli*

Introduction

Neuroscience has long characterized and categorized neocortex based on the functional properties that vary across its surface. Our understanding of the visual cortex, for example, largely derives from observations that the response properties of the ventral stream ascend in complexity. V1 is discussed as responding to "edge-detecting" Gabor filters [1], V2 to variations in local curvature [2], V4 to more complex shapes [3], and IT to specific objects and faces [4], which together have inspired the theory that object recognition proceeds via hierarchical image representations [5-7].

An area's response properties are most often characterized by varying stimuli while recording activity in that area. If neural activity changes robustly along a stimulus dimension, those neurons are sometimes said to 'encode' or be 'tuned' for that feature. This approach thus relies on building a response function, or tuning curve, along one dimension

of stimuli, or at most a small handful. However, natural stimuli are described by an enormously high number of dimensions. This means that there are necessarily many dimensions of stimuli that are left untested by any experiment that varies only a few dimensions of stimuli.

Leaving the response to dimensions of stimuli uncharacterized can complicate the interpretation of a tuning curve. In many studies it is hoped that tuning curves estimated from artificial stimuli will be good models of how neurons respond to stimuli in different contexts and thus represent their general functional role in visual processing. Most tuning curve experiments tacitly assume this to be true, at least approximately, as otherwise the interpretation would be unique and particular to the tested stimuli alone. The important question is how approximately tuning holds, and to what ranges of untested stimuli. For a tuning curve to be valid across multiple contexts the dimensions of stimuli varied in an experiment, which by necessity are few, must interact very little with the dimensions that were not characterized, which are likely very numerous. In other words, the neural response function must be able to be separated (additively or multiplicatively) into a term that depends on the varied dimension and a term that depends on uncharacterized dimensions. It is not clear that this is a reasonable starting assumption except in special circumstances, and ideally it should be tested how much tuning inferred from simplified stimuli is informative of tuning for more complicated stimuli.

It has been possible to examine the dependence of tuning on context only in some experimental paradigms in early visual areas. In V1, there has been a large effort to characterize neurons directly from their responses to natural images [8-13]. Some aspects of the V1 response appear the same the same for both natural images and drifting gratings, such as their preferred orientation [11]. However, other aspects of the V1 response are different for natural images [10], and this limits how well artificial stimuli responses can inform researchers about natural stimuli responses [12, 14]. For higher areas like V4, however, such a comparison has not been possible. The natural scene approach popularized in V1 requires approximating the response with a linear or second-order function of the image, but this is cannot be done accurately for V4. The nonlinearity of the V4 response is evidenced by the nonlinearity of the models that best predict V4 activity [5, 15, 16], as well as the fact that receptive fields (RFs) estimated from simple stimuli fail to predict much of the response to more complicated or natural stimuli [17-19]. Without access to interpretable receptive fields estimated with natural stimuli, it has not been possible to verify that experiments with artificial stimuli sets describe how V4 responds to natural images.

Tuning curve experiments are nevertheless put forward as a core component of our knowledge about the function and anatomy of V4. The study of its response to color has been particularly influential. V4 was first characterized as a color area [20] before later studies found selectivity for other visual features (such as orientation [21], curvature [3], shape [17, 22, 23], depth [24-26], and motion [27]; reviewed in [28]). The selectivity for different visual features is spatially clustered within V4. Color-selective neurons are predominantly located in color 'globs' [29], which intersperse 'interglob' regions more selective for orientation [30]. Glob cells are further arranged by their color preference [31] and generally in the same hue sequence as is found in perceptual color space [32, 33]. These findings have led to a hypothesis that V4 is anatomically segregated by function. It is important to note, however, that these findings largely follow from experiments in which the stimuli were colored gratings, oriented colored bars, or other single shapes. It is not known how accurately these tuning curves describe V4's response to color in natural stimuli.

In this work, we aimed to test the limits of how much tuning in V4 accurately generalizes to complex stimuli. We developed a new method to estimate the hue tuning curves of color-responsive neurons in macaque V4 from their responses to natural scenes, and then compared that to tuning estimated by varying the hue of simple stimuli. That is, we asked how well $P(Y|X, Z=z)$, which is the probability of spike counts Y given hue X and a fixed context z, stands in for $P(Y|X)$, the average hue tuning over natural images. Our method involved determining how hue affects the response of a general nonlinear model of the V4 response, which in our case was based on a deep artificial network pretrained to classify images [5]. We found that the tuning curves estimated from responses to stimuli of a uniform hue poorly described how hue affected responses to natural scenes. Thus, $P(Y|X, Z=z) \neq P(Y|X)$. Although hue strongly modulates the V4 response, hue tuning curves do not generalize from artificial settings to natural stimuli.

Results

We recorded the spike rates of neurons in area V4 of two macaques as they viewed images on a monitor. One monkey (M1) freely viewed images as we tracked its gaze, while the gaze of the second monkey (M2) was fixed at image center during image presentation. We analyzed the responses of 90 neurons in M1 over several viewing sessions, taking care that the identity of cells on each electrode did not drift across sessions (see Methods: Session Concatenation), and in M2 recorded from 80 neurons in a single session. We then estimated tuning curves from responses to both artificial and naturalistic stimuli in order to ask if and how hue tuning generalizes.

Tuning to hue on uniform screens

We first measured hue tuning by varying the hue of a uniform flat screen (Fig. 1A). We found that many of our neurons were well-tuned to specific hues (see examples in Fig. 1B), consistent with the previous literature on hue tuning in V4 [29, 30, 32]. We could consistently estimate hue tuning trials for 79/90 of neurons in M1 (Fig. 1C), but only for 17/80 neurons in M2 (Supp. Fig. 2A). A general trend across analyses was that neurons in M2 were more poorly described by hue than the neurons in M1. This difference in monkeys was possibly due to the spatial heterogeneity of color responses in V4 [29, 30]. In later analyses, we compared the hue tuning of neurons only when we could reliably estimate tuning.

Next, we asked if the tuning curves of the uniform hue context could predict natural scene responses. The V4 response is complex, but if the uniform field tuning curves accurately represent the contribution of hue, and the hue response is any considerable proportion of the overall response, then they should capture at least some variance. For example, we might expect that if a neuron preferred uniform fields of orange hue (like the examples in Figure 1), then that neuron would prefer scenes containing predominantly orange hues. Instead, we found that the images that elicited the highest spike rates were often composed of consistently different hues (Fig. 1D vs. Fig. 1E). We observed that the discrepancy between uniform hue tuning and natural scene responses was consistent across all neurons and trials (Fig 1F). Specifically, we asked how well uniform hue tuning curves could predict natural scene responses by interpreting the curves as the coefficients of a linear response to hue, and then scoring this model (see Methods). The pseudo-R^2 score of the tuning curve model was below zero for all but one neuron (Fig. 1F and Supp. Fig. 2C), which implies that variance predicted by the uniform field tuning curves gave worse predictions than the mean firing rate on natural scenes. Knowledge of V4 responses to single hues thus does not help to predict responses on natural images.

These observations can be explained if tuning shifts between contexts, or alternatively if these neurons respond more strongly to non-hue features that potentially co-vary within natural images. To distinguish these two possibilities, we next estimated tuning to hue from the responses to natural images and compared it with uniform hue tuning.

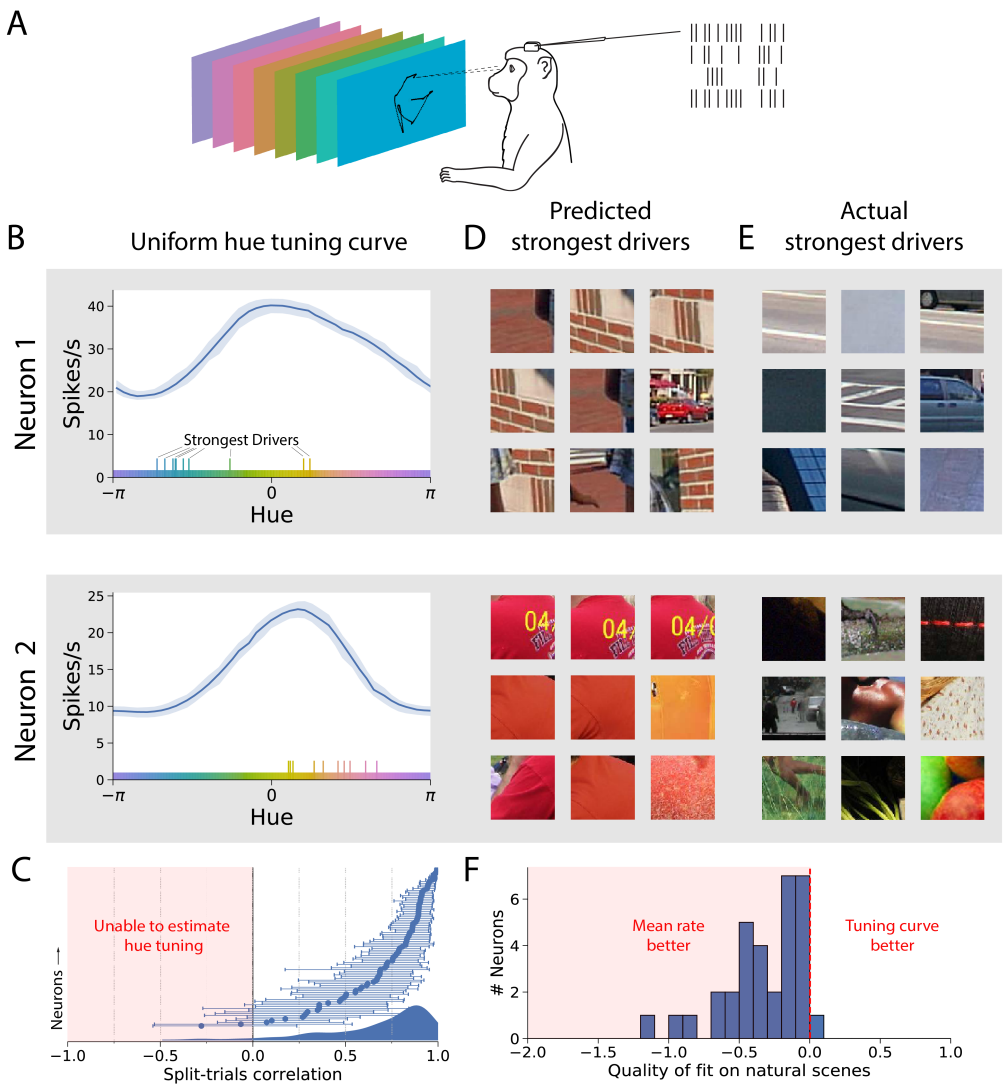

Figure 1. Tuning curves estimated from responses to artificial stimuli. Data from M1; see Supp. Fig. 2 for M2. A) We recorded from neurons in area V4 as a monkey viewed fields of a uniform hue. B) The uniform hue tuning curves for two example neurons, showing strong hue modulation. C) Our ability to estimate hue tuning for each neuron was captured by the correlation of the tuning curve estimated on two non-overlapping halves of the trials. This correlation would be 1 in the no-noise or infinite-data condition. D) Using the uniform hue tuning curve as a model of V4 activity on natural scenes, we would have expected these 9 trials to elicit the strongest responses. Each image displays only the image portion within the fixation-centered receptive field. E) In reality, the highest spike rates were observed on these 9 fixations. Neuron 1's strongest drivers were dissimilar in hue from the peak of the tuning curve, while those of Neuron 2 were somewhat consistent. The mean hue of each image (weighted by saturation) is shown as a tick in panel B. F) The natural scene responses on all trials and neurons were different than would be expected from the uniform hue tuning curves. Displayed here is the histogram of the Poisson pseudo-

R^2 goodness-of-fit scores of the tuning curves' predictions, which is below zero when the predictions underperform the mean firing rate.

Tuning to hue on natural scenes

In the context of natural images, one straightforward way to estimate hue tuning is to fit a (generalized) linear model to the natural scene responses using hues as covariates. In this approach, a tuning curve represents the mean change in the (log) firing rate observed with changes in each hue. This was our first and most simple method. A limitation of this model, however, is that it does not control for other visual features that drive V4 neurons, including interactions between hues. These factors will influence the hue tuning curve to the extent that hues and other features co-vary in natural scenes. To control for interaction effects, we additionally estimated tuning curves with two more complex models that each accounted for a greater number of possible drivers of V4. This progression was designed to investigate whether visual confounds could explain the discrepancy between uniform field hue tuning and natural scene hue tuning.

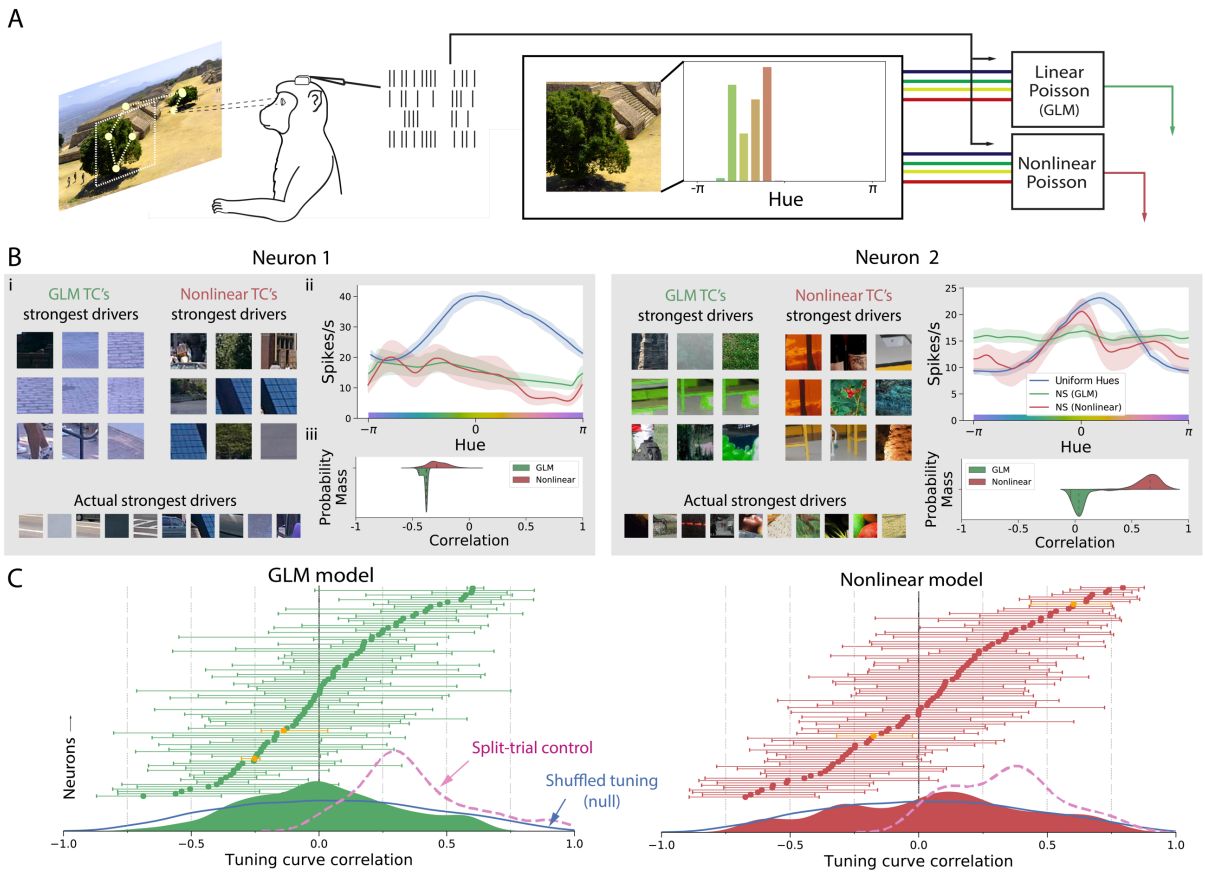

Figure 2. Tuning curves for hue estimated from the responses to natural images. Data from M1; see Supp. Fig. 2 for M2. A) We trained two models, a generalized linear model (GLM) with Poisson output and a nonlinear machine learning model (gradient boosted trees) with Poisson output, to predict each neuron's response from the hues present in its receptive field. B) (i) The 9 trials that each model predicted to have highest firing rate looked similar to trials with the actual strongest response, unlike the uniform hue model. (ii) We built tuning curves from each model. The uncertainty of each curve is given by the 5th and 95th percentiles of hundreds of model fits to the trials resampled with replacement. (iii) This uncertainty is then propagated into the correlation between the uniform hue tuning curves and natural scene tuning curves. C) The correlations of the natural scene and the uniform hue tuning curves across all neurons show that the natural scene and uniform hue hues rarely correlate. Above: The neurons are sorted by their correlation to show a cumulative distribution. The two example neurons are highlighted in orange. The error bars on each neuron show the 5th and 95th percentiles of the distribution of correlations observed while bootstrapping over model fits. Below: The smoothed density of all neurons' natural scene/uniform hue correlations is similar to what would be expected if neurons randomly shuffled hue tuning between conditions (overlaid, blue). Also overlaid (in pink) is the control distribution, which shows how well tuning estimated from one half of the natural scene trials correlated with the

tuning estimated on the other half. If hue tuning were the same across stimuli, then the distributions would look like the control distributions.

Generalized linear model of hue responses

We first modeled natural responses with a generalized linear model (GLM) upon hues (Fig. 2A). The covariates, detailed in Methods, are more specifically a histogram of the hues present in the neuron's receptive field on each fixation, with the contribution of each pixel weighted by its saturation. In monkey M2, the GLM could explain little activity (Supp. Fig. 3C), which prevented any analysis of hue tuning. (We were able to meaningfully estimate hue tuning in M2 only with our third and most complex model, below.) In monkey M1, the GLM successfully explained variance within the response to held-out natural scenes (Supp. Fig. 3B). This was a large improvement from the model estimated from responses to uniform fields. Indeed, compared to the predictions of the uniform hue tuning curves, the 9 strongest drivers of the model appeared more similar to the actual 9 strongest drivers (Fig. 2Bi). Thus, in M1, a hue model fit directly to natural scene responses provided better predictions than one fit to uniform hue responses.

The hue tuning of the GLM was measured from its responses to single hues, which is equivalent to inspecting the weights upon each hue (Fig. 2Bii, green curve). To see if hue tuning changed between natural scenes and uniform hues, we correlated the tuning curves across contexts (Fig. 2Biii). Across all neurons analyzed in M1 (Fig. 2C), the correlation between the tuning curves of the two stimuli sets varied widely and had a mean not significantly different from zero (mean of 0.01 [-0.06 0.06], 95% bootstrapped confidence interval). In fact, the spread of correlations was similar to the distribution that would arise by chance if hue tuning changed randomly between contexts (Fig. 2C inset), which we approximated by shuffling the neurons and cross-correlating their uniform field tuning curves. Thus, the tuning curves of the GLM hue model fit to natural scenes were quite dissimilar from the uniform field tuning curves.

Since the correlation across contexts would also appear to be lower due to noise or simply a bad model fit, it was important to quantify our uncertainty of the tuning estimation. We repeatedly refit the GLM on the natural scene trials resampled with replacement, and observed the distribution of coefficients. This distribution was propagated through the analysis to obtain a distribution of curve correlations (Fig. 2Biii) whose 95th percentiles form the confidence interval of the natural scene (NS)/uniform field correlation for each neuron. Since correlations of exactly 1 would be impossible in the presence of any sources of noise in curve estimation, we also visualized how high correlations would have appeared if tuning were the same in both contexts, given all sources of noise. This was estimated by comparing

hue tuning on two non-overlapping halves of natural scene trials (Supp. Fig. 1A). The correlations between natural scene and uniform hue tuning were significantly lower than this control, ($p=3.2x10^{-12}$, Wilcoxon signed-rank test; see Fig. 2C for the population distributions and Supp. Fig. 1D for the per-neuron comparison). Note that the split-trial control is a conservative lower bound of our quality of estimation, as the model was fit on only half the number of trials. Thus, uncertainty in our curve estimation cannot explain away our observation of a difference in the hue tuning of the GLM fit to natural scenes and uniform hue tuning.

Nonlinear model of hue responses

A shortcoming of the GLM is that it does not model interactions between hues. Nonlinear hue interactions have been previously observed in V4 [34]. This would lead to a bias in the GLM's hue tuning because hues are correlated in natural scenes (Supp. Fig. 3A). To test if this bias could explain the observed difference in hue tuning, we fit a second model that included nonlinear interactions between hues. We fit a machine learning model (gradient boosted decision trees, via XGBoost) to predict the neural response from the histogram of hues present in each natural image fixation. This model, which we refer to as the 'nonlinear hue model', predicted neural activity more accurately than the generalized linear hue model for all neurons in both M1 and M2 (Supp. Fig. 3B,C). This confirmed that these neurons responded nonlinearly to hue. It is important to note that because of this nonlinearity, no one-dimensional tuning curve could represent the full hue response. It would be necessary to estimate multi-dimensional hue tuning curves to display interactions between hue bins. Our focus here is instead on the average response to individual hues on natural scenes, and whether this average hue response was similar to hue tuning on uniform hues.

We estimated hue tuning curves for the nonlinear hue model fit on natural scene responses by measuring its responses to single hues, in essence reproducing the uniform hue experiment but on the natural scene model. In neurons for which we could consistently estimate hue tuning in this manner, we found that these tuning curves correlated poorly with the uniform field tuning curves (Fig. 2B,C). However, they correlated strongly with those estimated from the GLM (Supp. Fig. 4A), indicating the bias due to nonlinearity and hue correlations was small. As we did for the GLM, we estimated the uncertainty of this model by correlating tuning curves estimated on non-overlapping halves of data. The nonlinear hue model was able to consistently estimate hue tuning for many neurons in M1 (Fig. 2C overlay) but for just two neurons in M2 (Supp. Fig. 2 D-F), which prevented a statistical analysis in M2. In M1, the natural scene/uniform field tuning curve correlations were significantly lower than these split-trial correlations ($p=1.0x10^{-14}$,

Wilcoxon signed-rank test; Supp. Fig. 1D), indicating that the observed change in hue tuning across contexts was not a consequence of noise in the estimation of tuning. Thus, after accounting for interaction effects between hues, we found that uniform field tuning does not generalize to estimated natural scene tuning.

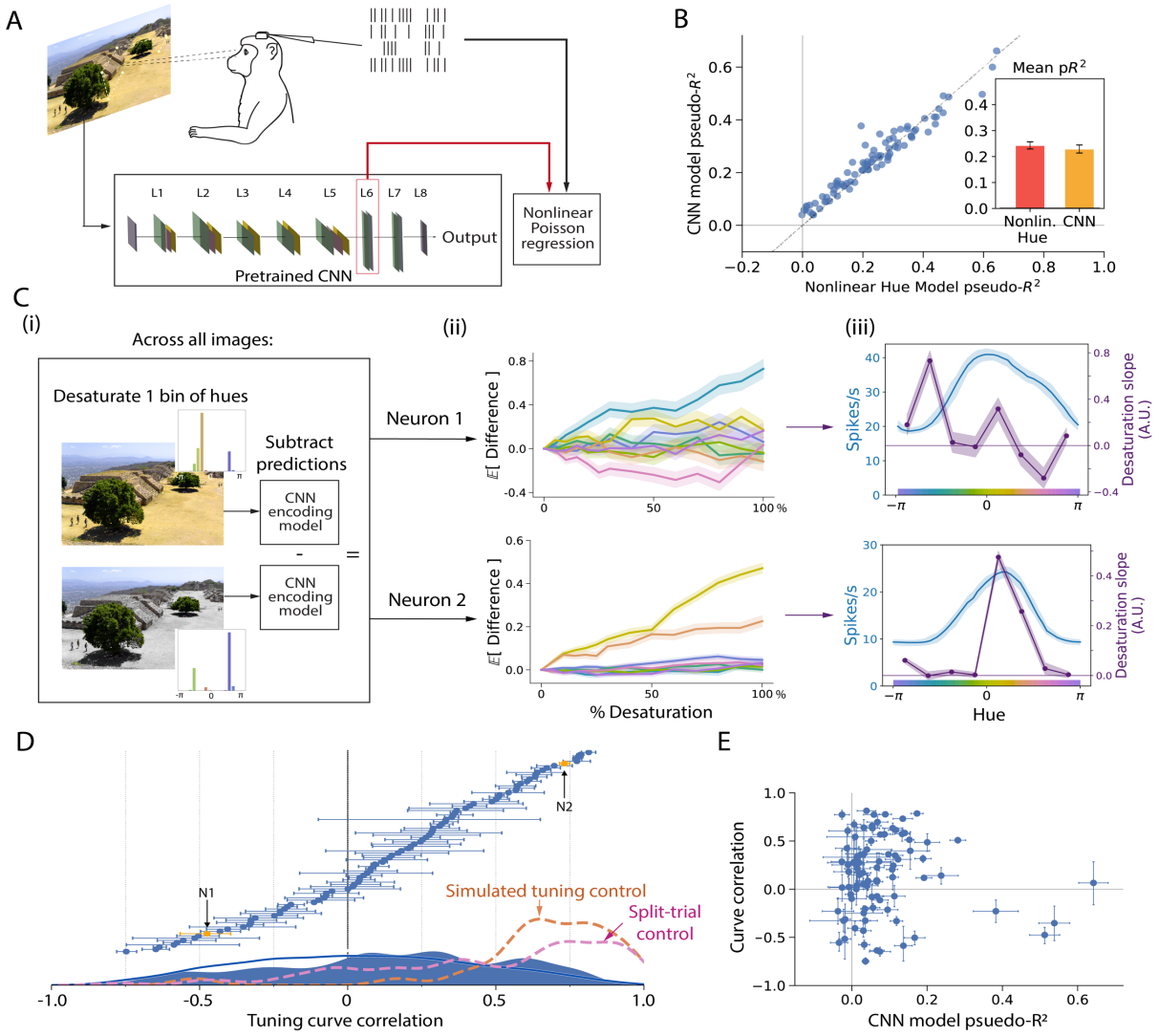

Figure 3. Tuning curves estimated for hue from a model of V4 responses built from a pretrained convolutional neural network (CNN). Data from M1; see Supp. Fig. 2 for M2. A) We trained a nonlinear Poisson regression model (gradient boosted trees) to predict the V4 response from the activations of an intermediate layer in the VGG16 network given the visual stimulus. B) The quality of the neural predictions on each neuron, measured by the cross-validated pseudo-R^2 score, were similar between the CNN model and the nonlinear hue model. C) We built hue tuning curves in the following manner: (i) For each image in a test set, we slightly desaturated all pixels in a bin of hues, and subtracted the CNN model's predictions on the perturbed image

from those on the original image. (ii) For each neuron, the average change in the predicted response across all test images was plotted against the percentage by which hues were desaturated. The slope of each line is, to first order, the average effect of that hue on the model response in the test set. The top and bottom plots show the same example neurons as in earlier plots. (iii) The resulting tuning curve (purple) summarizes the average effect of each of the 8 bins of hues – i.e. the slopes of the 8 desaturation curves. It can be seen that the tuning of neuron 1 was poorly correlated with the uniform hue tuning (blue), while that of neuron 2 was well-correlated, in agreement with the hues of the strongest-driving stimuli shown in Fig. 1B. D) We calculated the correlation between the two tuning curves for all neurons. The distribution of correlations was lower than for the reconstructed hue tuning of simulated neurons ("simulated tuning control"; see also Supp. Fig. 5) as well as the distribution of correlations between tuning curves estimated from two non-overlapping halves of the natural scene trials ("split-trial control"; see also Supp. Fig. 1). E) The quality of the CNN model fit for each neuron did not predict the correlation of the tuning curves.

180 ***Neural network model of V4 responses***

181 We next repeated the estimation of hue tuning on natural scenes with a more general model of V4 neurons that does
182 not rely on hand-specified summaries of the features present in a receptive field. This was important to ensure that our
183 results were not sensitive to design decisions in processing the images, as well as to account for the confounds of
184 other, non-hue features contained in the image. The two hue models would provide biased estimates of tuning if
185 neurons also responded to other visual features, and if these features co-varied in the image dataset with hue. If most
186 green objects are plants, for example, the observed dependence on green hues may be partially attributable to a
187 response to the high spatial frequency of greenery. Theoretically, one could include these features as additional
188 covariates, but the list of features that drive the V4 response in a simple manner (e.g. linearly) is not comprehensively
189 known. Good progress has been made with shape and texture [3, 35, 36], but arguably not enough to thoroughly
190 control for all non-hue features in a simple model. Controlling for other drivers of V4 thus requires a model that learns
191 relevant visual features instead of using features chosen by a researcher or parameterized by hand.

192 The model we selected was based on a recent encoding model of V4 that relates neural responses to the activations to
193 the upper layers of a convolutional neural network (CNN) pretrained to classify images [5]. Such "transfer learning"
194 models have also recently been used to infer optimal stimuli for neurons in V4 [15, 16]. Instead of pre-specifying a
195 receptive field estimated with sparse noise, we allowed the CNN model to learn any location sensitivity itself and thus
196 fed the entire fixation-centered image as input. Predictions of neural activity are obtained by passing the image through
197 the network, obtaining the intermediate network activations, and then passing these to a classifier trained for each

neuron (Fig. 3A). The predictions of neural activity given by this model were comparable in accuracy to those of the nonlinear hue model (Fig. 3B) despite the model making many fewer assumptions about how raw pixels related to responses.

Our initial, unsuccessful method to estimate hue tuning from this model was to simply observe the model's response to images of a uniform hue, as before. However, this approach failed to reconstruct tuning on simulated data. This interesting parallel to our main finding is likely due to feature interactions in the model and the fact that uniform field test images are far outside the domain of natural scenes on which the CNN was pretrained.

Instead, we developed a method to estimate hue tuning from the model that only uses responses to images close to the domain of natural images. The method requires slightly perturbing the hue of input images and observing the change in the learned model's response, thus testing to which hues the model is most sensitive in natural images (Fig. 3C). First, for a test set of images not used for training, we desaturated all pixels within a bin of hues by a set percentage (Fig. 3Ci). The percentage of desaturation varied from 0% (i.e. no change) to 100% (in which all pixels of one hue are taken to isoluminant grey). We took the difference between the model's predictions on the original and perturbed images and examined how severely this difference depended on the level of desaturation (Fig. 3Cii, iii). For each neuron, we averaged over the entire image dataset to yield the average effect of perturbing each hue on natural images. This method established the effect of hue only in the tight neighborhood of each image, and is set up to estimate the average local effect of hue on the natural image response.

To ensure that this process could in principle reconstruct correct tuning curves, we built simulated responses (Supp. Fig. 5). We generated random cosine tuning curves, then simulated a hue response by applying these as linear filters upon the histograms of the hues present in each image. We then attempted to predict these simulated responses from the activations of the pretrained CNN given the raw images. Using the method of progressively desaturating test images, we found we could reconstruct the original cosine tuning curves with high accuracy (Fig. 3D overlay and Supp. Fig. 5), even though the pretrained CNN model was trained to classify images and not to extract hues. As a second, more conservative test, we also performed the split-trial control for the actual V4 neurons, which involved repeating the entire analysis separately on two non-overlapping halves of natural scene trials and then correlating the two resulting tuning curves. The split-trial tuning curves showed significantly positive correlations for most neurons

in M1 (Fig. 3D overlay) as well as for neurons in M2 (Supp. Fig. 1). This method of querying the effect of hue could thus accurately estimate hue tuning curves from natural scene responses in both monkeys.

We next asked if these tuning curves would be similar to the tuning curves to uniform hues. If hue affected V4 responses in the same way in both contexts, we would have observed the correlations between tuning curves across contexts to be at least as positive as the split-trial control. This was not the case. We found that the tuning curves of one context were different from tuning in the other (Fig. 3D for M1 and Supp. Fig. 2I for M2). Among those neurons for which we could consistently estimate hue tuning, the natural scene/ uniform hue tuning curve correlations were significantly closer to 0 ($p=1.1x10^{-8}$, Wilcoxon signed-rank test, Supp. Fig. 1D for M1; Supp. Fig. 2I for M2). This difference in tuning curves was not an artifact of our model fit or estimation method, as this would be measured in the split-trial control, and additionally we observed no correlation between the model's accuracy on unseen natural images and the natural scene/uniform field correlation (Fig. 3E and Supp. Fig. 2K). In addition to changes in tuning curve shape as captured by correlation, we also examined if the natural scene tuning curves showed changes in the overall degree of hue modulation. We found that hue modulation – the maximum of a tuning curve minus the minimum, normalized by the mean – was related across contexts, but weakly (Supp. Fig. 6). Many neurons strongly modulated by hue on uniform fields had weak responses to hue on natural scenes, and vice versa. Overall, the tuning curves estimated with this more advanced method support our previous conclusion that hue tuning on uniform fields does not agree with the effect of hue in natural scenes.

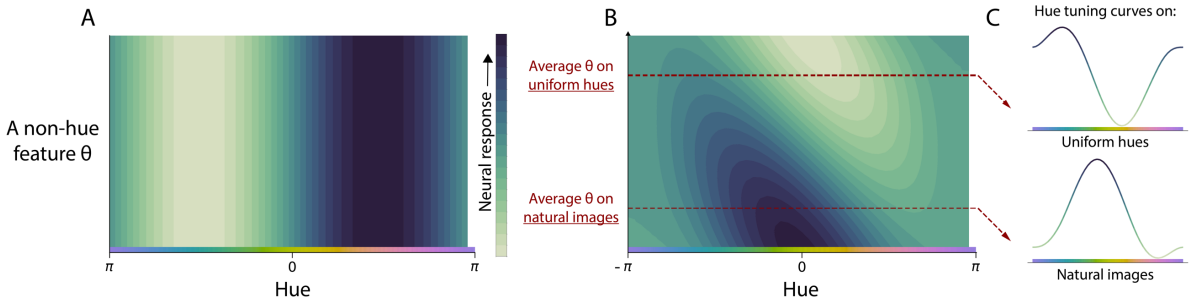

Figure 4: Interactions between features allow neurons to carry more information in their activity. A) In this two-dimensional tuning curve, a hypothetical neuron responds to only hue and carries no information about other variables. B) A hypothetical neuron that responds as well to another non-hue feature is informative about multiple dimensions of stimuli (due to its nonzero derivative). C) We can build a hue tuning curve for this neuron by varying hue with the other feature held fixed.

If the average non-hue feature is different between natural images and uniform hues, the tuning curves to hue will differ between contexts.

Why interactions between features?

A straightforward explanation of why hue tuning differs across visual contexts is that these neurons respond to nonlinear combinations between hue and non-hue features, as shown schematically in Figure 4. There must be a computational advantage that explains this coding scheme for visual perception. It is clear that if the role of these neurons were to encode hue alone, then any nonlinear interactions would be detrimental. This is because hue can no longer be unambiguously read out without additional contextual information. Therefore these V4 neurons likely assist in a more general task, like object recognition or segmentation. Other studies have also noted that color vision may be best thought of in terms of task performance [37]; the absorbance spectra of the L and M photoreceptors in primates, for example, are not maximally separated as in birds but rather overlap significantly, possibly because this helps to discriminate and classify fruit and leaves [38]. The question then arises: why would neurons being responsive to multiple features help visual processing?

A simple strategy that predicts nonlinear interactions is to minimize the error of any read-out of encoded information from V4 to other brain areas. We can make this idea precise by referring to the notion of Fisher information, which bounds the mean squared error of any optimal readout from population activity (see Supplementary Information for additional details). This framework pulls from a large literature relating to optimal coding strategies [39-41]. The Fisher information is higher – and the potential decoding error is lower – when the neural population activity is highly sensitive to changes in the task-relevant features. One way to increase the population sensitivity to the task (i.e. the Fisher information) is to have each neuron be sensitive to multiple features. This will increase the total number of neurons in the population that are sensitive to each feature; when each neuron responds to k features instead of just one, k times more neurons respond to each feature on average. By increasing this number, the Fisher information also increases (though see below), and the achievable performance on the task increases.

Eventually, however, further increasing the k number of features to which a neuron responds will deteriorate how precisely it can respond to other features, decreasing the Fisher information. Depending on neural physiology (for example, the maximum firing rate, synaptic noise levels, and correlated variability), this tradeoff determines the

optimal number of features that a neuron should respond to. It is an extreme and unlikely case when the optimal number is one. Indeed, several publications have found that in common scenarios, like linear-nonlinear responses [40] and von Mises tuning curves [42], feature interactions are usually optimal. In particular, some degree of interaction is always optimal when the features co-vary in the natural world. This is the case with hue and most visual descriptors, and so it could be expected that V4 neurons would show interactions between hue and other visual features.

DISCUSSION

For populations of V4 neurons in two macaques, we found that varying the hue of simple stimuli produced tuning curves that do not accurately describe hue tuning measured from natural scenes. While some discrepancy may be expected, we found that the two sets of tuning curves correlated not much better than chance. This finding was robust across multiple methods of estimating tuning, which together accounted for the confounds of both hue-hue interactions as well as of non-hue drivers of V4 activity. A hue tuning curve for V4 estimated from any one set of stimuli thus does not universally describe the average response to hue on other stimuli. This implies that V4 neurons, including those with strong hue modulation, respond to nonlinear combinations of hue and non-hue information.

Known sources of modulation in visual responses

The V4 response is modulated by a number of factors that change with visual context. These factors are divided in the manner of their relevance to our findings. First are possible reasons why we might have observed low tuning curve correlations even if, in fact, tuning did not change between contexts. The second category of factors are known interactions between hue and other features in the V4 response. These are possible explanations of why hue tuning in V4 changes with visual context. We will review both in turn.

Of first concern as a potential confound upon hue tuning estimation is visual attention [43]. A particularly relevant form of attention is feature-based attention, in which neurons tuned for a feature (say, red) increase their firing rate if that feature is attended to (as in the task, “find the red object”) [44, 45]. While the task of M1 was free viewing and involved no instructions, it is likely that the monkey’s attention shifted during the task and that it was influenced by object salience. This effect may bias our results if object salience were correlated with hue. Attention is less likely to have presented a confound in the task of M2, in which gaze was fixed at center and stimuli were presented for 100ms. We have not directly controlled for attention, apart from trends in salience that might have been learned by the CNN

model, but we believe that the size of the apparent change in hue tuning cannot be attributable to salience-hue correlations.

Neurons in V4 are have been shown to preferentially respond to objects near the center of attention, even when attention falls away from fixation [46-48]. This phenomenon of receptive-field remapping is most problematic for our GLM and nonlinear hue models, which required that we extract the hues lying within the receptive field. If the monkeys' attention frequently strayed away from fixation, we would have extracted hues from an irrelevant image portion. This would introduce some noise in the hue covariates, and therefore some smoothing of hue tuning curves. The CNN model learned any spatial sensitivity directly from the natural scene responses. However, the effect of attention upon receptive fields could not be modeled and it is likely that some smoothing of the hue tuning curve occurred for this technique as well. Smoothing would obscure fine-scale structure in the tuning curves. As the curves were already smooth, however, the natural scene/uniform field correlations should not be much diminished. The smoothing effect is furthermore not consistent with our finding that many neurons have natural scene hue tuning with zero, or even negative correlation with their uniform field tuning while still showing strong hue-dependent modulation.

We now turn to potential descriptions of the interactions that might have led to a shift in hue tuning across contexts. One possibility is the behavioral phenomenon of color constancy, which would present as a neural correlate as responses to the inferred surface color of objects rather than their apparent color (which reflects the color of ambient light) [34]. This is a clear example of the nonseparability of the V4 response to hue, and a reason hue tuning might change between any two, single stimuli. It is less obvious, however, that color constancy correlates would cause the average effect of hue over all natural images to be different than on uniform hues. It would be expected that over tens of thousands of images with a broad range of lighting conditions, color constancy would result in some smoothing of the estimated tuning curve due to the difference between the pixels' hue and the inferred hue, and of the same characteristic scale as their typical difference. Additionally we may expect a bias that would result from the discrepancy between pure white and the average lighting condition. We expect this discrepancy to be small, and therefore that natural scene tuning curves would still be strongly (though not perfectly) correlated with the uniform field tuning curves. Though phenomena like color constancy would affect hue tuning on natural scenes, it cannot account for the entire difference we observed, and it is likely that there exists other undocumented sources of nonseparability.

A subpopulation of neurons in V4, so-called equiluminance cells, respond to object boundaries defined solely by chromatic boundaries [49]. Such shapes are defined by changes in hue or saturation, and so it is worth asking whether the response function of equiluminance cells includes interactions between hue/saturation and spatial arrangement. However, it was not originally determined if the responses were actually separable in this way, as neurons' hue tuning curves were characterized with a fixed shape. It is possible that equiluminant cells had fixed hue tuning that was then modulated by shape. Thus, it is plausible but undetermined that equiluminance cells would show different hue tuning across shape and explain our results.

The apparent shift in hue tuning in natural scenes may be partially be explained by a multiplicative or generally nonlinear interaction between shape and color, as was examined in a recent paper that jointly varied the hue and shape of simple stimuli [50]. In a linear model with terms for hue and a multiplicative interaction between hue and shape, the authors observed a significant interaction between shape and color in the majority of cells (51/60) as well as a color-dependent gain on shape tuning (44/60 cells). This interaction would cause (linear) hue tuning to appear different for natural images with varying shapes, as we observe. We note that other, undescribed features may also interact with hue, and that conclusively determining which visual features interact would require presenting stimuli tiling many more dimensions of variation.

Implications for V4 and for the tuning curve approach

Color responsivity has long been a defining feature of V4 [20, 51]. Recent studies have shown that localized areas in V4 are strongly responsive to color [29], and furthermore that the anatomical organization of color preference on the neocortex is similar to perceptual color spaces [31-33]. These findings have been taken as evidence that areas within V4 are specialized for the perception of color. However, each of these studies characterized hue tuning by changing the color of simple shapes. Since the color tuning of V4 neurons changes with visual context, as we show here, it is possible that previous conclusions about the functional organization of V4 do not accurately describe how V4 processes more naturalistic stimuli.

Some previous studies, based on the discovery of robust tuning for the color of simple visual stimuli, have concluded that the role of color-responsive areas in V4 is to represent color. Our results do not rule this out; for example these areas might represent color but be modulated by what colors are likely given the surroundings. This would complicate a read-out of color from V4, but may have other advantages like efficiency. It would be interesting to investigate this

possibility in future studies. An alternative possibility is that the color-sensitive areas of V4 are not specialized to represent color, *per se*, but rather serve a more complex role within recognition and perception. This is analogous to how V2 appears tuned to orientation but can perhaps be better described as processing naturalistic texture [52] . Furthermore, this role aligns with the suggestion that the ventral temporal cortex at large decomposes scenes into neural activity such that object categories are linearly separable [53]. Thus, the color-responsive areas of V4 may represent how color informs an inference of object identity [37]. Whether the color responses of V4 are an end to themselves (i.e. representing color) or intermediate computations in a larger assessment of object identity [54], or both, cannot be decided from this study; both are consistent with the data.

Our study joins a longer history of literature observing that, across many brain areas, tuning curves previously characterized with simple stimuli in fact change with context. In V1, for example, researchers found that receptive fields change with certain visual aspects that were not varied within previous stimuli sets, such as the presence of competing orientations [55][56-58]. Even sound has been shown to modulate V1 receptive fields, at least in mice [59]. More recently, it was observed that receptive fields are different in the contexts of dense versus sparse noise for neurons in layer 2/3 of V1 [60]. Spatio-temporal receptive fields of V1 neurons also appear different when estimated on natural movies versus drifting gratings [10, 12] (though note that orientation tuning is similar for static natural scenes versus gratings [11]). In other areas, such as for retinal ganglion cells [61, 62] and in macaque M1, S1, and rat hippocampus [63], contextual modulation in the form of nonlinear feature interactions have been identified by comparing the performance of a model that assumes separability (such as a GLM) with a nonlinear model that does not. Thus, while tuning curves generalize in some situations (e.g. [11]), it is common that they do not, and any assumption of separability of the neural response should be verified. Furthermore, as discussed in Results and the Supplementary Information, feature interactions are likely optimal for visual processing when the full visual scene is represented in neural activity and should be expected. Unless specifically investigated, it might not be correct to assume that a tuning curve accurately describes the neural response on different stimuli than used to create it.

If it cannot be assumed that neural tuning is separable, however, it becomes necessary to test prohibitively many stimuli or else make an alternative simplifying assumption. This is because the stimuli must scatter the entire space of relevant features, rather than be systematically varied along just one feature at a time. Since the number of tested stimuli must follow the number of potential feature combinations, the overall number of stimuli will grow

exponentially with the number of features. When there are very many features, even very large recording datasets by today's standards may be insufficient.

One possible way forward is to make simplifying assumptions, i.e. to set strong priors of the kinds of tuning curves that could be expected. This is the approach taken, for example, when modeling neurons using the activations of deep neural networks pre-trained on image classification tasks [5, 64] or considering neural responses as implementing a sparse code [9, 65]. To compare with the previous literature, single dimension experiments can then be performed on these complex encoding models, as we demonstrate here, or alternatively performed directly on artificial neural networks to gain intuition about what tuning curves say about information processing [66, 67]. In general, finding suitable priors will require the use of strong theoretical ideas and mechanistic hypotheses. To estimate tuning without assuming separability, then, neurophysiology must embrace and develop theories of neural processing.

Materials and Methods

Experimental setup: recordings

In each of two male macaques, we recorded from 96-electrode Utah arrays (1.0 mm electrode length) implanted in visual area V4. At the time of the experiment, Monkey 1 (M1) was aged 5 years, 10 months and Monkey 2 (M2) was aged 9 years, 4 months. Surgical details describing the implantation method can be found in previous publications [68, 69]. The array was located in the left hemisphere for monkey M1 and in the right hemisphere for M2. Spikes were sorted off-line first with an automated clustering procedure [70] and then refined by hand using custom MATLAB software (https://github.com/smithlabvision/spikesort) taking into account waveform shape and interspike interval distributions [71].

All experimental procedures were approved by the Institutional Animal Care and Use Committee of the University of Pittsburgh.

Gaze tracking and fixation segmentation

We employed a free-viewing paradigm for one monkey (M1) and a fixed-gaze paradigm for the other (M2). The location of each monkey's gaze on the screen was tracked with an Eyelink 1000 infrared tracker (SR Research, Ottawa, Ontario, Canada). Visual stimuli were presented and the experimental trials were controlled by custom Matlab

software in conjunction with the Psychophysics Toolbox [72]. For monkey M1, we segmented each fixation as a separate event based on thresholding the position and velocity of the gaze coordinates. We did not analyze activity occurring during eye movements. Once each fixation was separated, the average location of the fixation was recorded and matched to image coordinates. Monkey M2 was trained to fixate on a dot positioned at the center of each image. The gaze was tracked as for M1, but this time only to enforce fixation and terminate the trial if the gaze shifted away from center.

Artificial stimuli

Both monkeys viewed uniform images of a single hue on a computer screen at 36 cm distance, with a resolution of 1024x768 pixels and a refresh rate of 100 Hz on a 21" cathode ray tube display. The full monitor subtended 55.5 degrees of visual angle horizontally and 43.1 degrees vertically. The monitor was calibrated to linearize the relationship between input luminance and output voltage using a lookup table. This calibration was performed for grayscale images, and the color profile of the monitor was not separately calibrated. The hues were sampled from the hue wheel in CIELUV color space (calculated with a D65 standard illuminant and standard observer) at increments of 1 degree, and were presented in random sequence. Monkey M1 freely viewed the stimuli, and was rewarded periodically for maintaining eye position on the screen for 4 seconds, after which time the static image was refreshed. The trial was ended if the monkey looked beyond the screen during this duration. Monkey M2 was trained to fixate a small dot at the center of the screen for 0.3 seconds, during which three images were flashed for 100ms each. A 0.5 second blank period interspersed each fixation. Monkey 1 viewed 7,173 samples of the uniform hue stimuli over 10 sessions, while Monkey 2 viewed 1,119 samples during a single session.

Natural images

Both monkeys viewed samples from a dataset of 551 natural images, obtained from a custom-made Google Images web crawler that searched and downloaded images based on keywords such as cities, animals, birds, buildings, sports, etc. Monkey M1 viewed images over 15 separate sessions, for a total of 77961 fixations. Monkey M2 viewed images over two sessions on a single day, for a total of 6713 fixations. We then extracted the features from the image patch centered around each fixation that would serve as model inputs. The image patch around fixation corresponded to the 400 x 400 pixel block surrounding the center of gaze. This corresponds to a region 23.5 visual degrees on a side.

For the nonlinear methods we included a small number of features unrelated to the images as additional controls. To account for possible stimulus adaption, we included the trial number in the session and also the number of times the monkey previously fixated on that image. While all models predict the spike rate, which is already normalized by the fixation duration, we included the fixation duration as an input to control for possible nonlinearities of rate with fixation duration. We also included the duration of the saccade previous to the current fixation, the duration of the saccade after fixation, the maximum displacement of the gaze position during the entire duration of the fixation, and whether the pupil tracking was lost (often due to a blink) in the saccade before or after fixation. Including these inputs allowed the nonlinear methods to control for factors which also may affect spike rate.

Receptive field estimation

To estimate hue tuning on natural scenes with the hue models, we needed to know which hues were present within the RF on each fixation. We mapped the RFs by presenting sinusoidal gratings at four orientations, which were flashed sequentially at the vertices of a lattice covering a portion of the visual field suggested by anatomical location of the implant. For monkey 1, this procedure identified an RF of the implant of 5.87° in diameter (full-width at half-maximum) centered 8.94° below and 4.99° to the right of fixation, whereas for M2 we found an RF 7.02° wide centered 7.02° below and 7.02° to the left of fixation. The location of the RF was confirmed in the natural scene presentations as the pixel block that allowed the best predictions on held-out trials. For the hue model analyses, on each fixation we obtained the model inputs by extracting the hues present in the 50x50 pixel block (2.93º of visual angle on a side) surrounding the centroid of the RFs of each monkey.

We did not use this RF information in the CNN model, which took as input the entire image region around the fixation. Since information about spatial location preserved in the lower and intermediate layers of the CNN, the RF for any neuron can be learned. This addressed any worry that the RF specification might systematically change for natural images.

Session concatenation

Although all recordings in M1 were performed with the same implanted Utah array, they were recorded over several sessions. The recordings for M2 were made in a single session. In M1, this introduced the possibility that the array might have drifted, and that a single channel might have recorded separate neurons in different sessions. To address this possibility, we noted that spikes identified in a channel in one session will be less predictive of another session's

activity if the neurons are not the same, as we expect tuning to be relatively static across days [73, 74]. We thus filtered out neurons whose uniform hue tuning changed across sessions. We trained a gradient boosting regression model with Poisson targets to predict spike counts in response to the hue of the stimuli. Nuisance parameters, such as duration of stimulus, gaze position, inter-trial interval, etc., were also included as model covariates to increase the predictive power even for neurons that were not hue-tuned. We then labeled a neuron as having static tuning as follows. First, we trained the model on each single session in a 8-fold cross-validation procedure and recorded the mean pseudo-R^2 score. This score reflected how well the model could predict held-out trials on the same session. Then, we re-trained the model on each session and predicted on a different session, for all pairs of sessions. This resulted in a cross-prediction matrix with diagonal terms representing same session predictability (the 8-fold CV score), and off-diagonal terms representing generalization between sessions. We did not concatenate sessions if hue tuning estimated in one session could not predict hue responses in another session (i.e. the CV pseudo-R^2 score was less than 0).

The natural image sessions were interspersed with the artificial sessions. If a natural image session occurred between two artificial sessions, and a neuron showed static tuning both artificial sessions as identified in the above manner, then that natural image session was included for the hue tuning comparison and model fitting. The recordings of units from other natural image sessions were not used. This procedure improved our confidence that the neurons recorded in different sessions were the same.

Uniform hue tuning curve estimation

Hue tuning curves were built for each neuron by plotting its spike rate on each fixation against the observed hue. For the visualizations in the figures, we performed LOWESS smoothing, in which each point of the curve is given by a locally-weighted linear regression model of a fraction of the data. The error envelope of the curve represents the 95% confidence interval given by bootstrapping over individual fixations. To calculate the correlation between tuning curves, we did not correlate the LOWESS-smoothed curves but rather the simple binned averages. We created 16 bins of hues and calculated the average spike rate for all stimulus presentations of those hues, then correlated the 16-dimensional tuning curve vector with the natural image tuning curves.

Natural scene models

Model scoring and cross validation:

We quantified how well the regression methods described neural responses by calculating the pseudo-R^2 score. This scoring function is applicable to Poisson processes, unlike a standard R^2 score [75]. The pseudo-R^2 was calculated in terms of the log likelihood of the true neural activity $L(y)$, the log likelihood of the predicted output $L(\hat{y})$, and the log likelihood of the data under the mean firing rate $L(\bar{y})$.

$$R^2 = 1 - \frac{\log L(y) - \log L(\hat{y})}{\log L(y) - \log L(\bar{y})} = \frac{\log L(\hat{y}) - \log L(\bar{y})}{\log L(y) - \log L(\bar{y})}$$

The pseudo-R^2 is, at left, one minus the ratio of the deviance of the tested model to the deviance of the null model. It can be also be seen, at right, as the fraction of the maximum potential log-likelihood. It takes a value of 0 when the data is as likely under the tested model as the mean rate, and a value of 1 when the tested model perfectly describes the data.

We used 8-fold cross-validation (CV) when assigning a final score to the models. The input and spike data were segmented randomly by fixation into eight equal partitions. The methods were trained on seven partitions and tested on the eighth, and this was repeated until all segments served as the test partition once. We report the mean of the eight scores. If the monkey fixated on a single image more than once, all fixations were placed into the same partition. This ensures that the test set contains only images that were not used to train the model.

Hue models

The uniform field linear model, the generalized linear hue model, and the nonlinear hue model all describe neural activity as a function of the hues present in the receptive field on each fixation. To build the histograms, we calculated the hue angle of each pixel in CIELUV space, and then calculated the number of pixels in each of 16 bins of hues. Note that a hue is defined for a pixel even if it is quite desaturated. To ensure near-gray pixels would not affect the results, we weighted the contribution of each pixel to the histogram by its saturation (defined as the distance of the color from the L axis). Since the hue histograms have 16 bins, the base regression problem to describe neural activity from hue is 16-dimensional.

The uniform field model, presented in Figure 1F, is a linear model whose coefficients are set from the uniform field tuning curve. Inference is performed via a dot product of the coefficients with the hue histogram. This is, we multiplied the mean firing rate observed for a bin of hues by how much that hue bin is present in the receptive field, and then summed across hue bin. We then added a constant term to account for the difference in mean firing rate across contexts.

The generalized linear model (GLM) was a linear-nonlinear model with an exponential link function and a Poisson loss. We included elastic net regularization, and selected the regularization coefficient for each neuron using cross-validation in an inner loop. We implemented this with the R package r-glmnet [76]. For our nonlinear model, we selected the machine learning method of gradient boosted decision trees as implemented by XGBoost, an open-source Python package [77]. This method allows a Poisson loss function and has previously been shown to be effective in describing neural responses [63]. Briefly, XGBoost trains multiple decision trees in sequence, with each trained on the errors of the previous trees. We chose several regularization parameters using Bayesian optimization for a single neuron. These parameters included the number of trees to train (200), the maximum depth of each decision tree (3), the data subsampling ratio (0.5), the minimum gain (0.3), and the learning rate (0.08).

To build tuning curves from the fit GLM and XGBoost models, we predicted the response to a vector indicating which color was present (that is, a "one-hot" vector with one entry per hue that is all zeros except for the hue that is present). Then, to estimate the measurement error of the tuning curves, we refit the models to the original neural responses resampled with replacement. This resulted in tuning curves from hundreds of bootstrapped model fits. In figures in which we display the tuning curves, the lower and upper error bounds represent the 5th and 95th percentiles of the tuning curves observed when refitting the models to the resampled data.

CNN model

Our convolutional neural network (CNN) encoding model was based on previously published studies in which it was shown that the intermediate layers of pretrained networks are highly predictive of V4 responses [5]. Ours was built from the VGG16 network, which is a large convolutional network trained to classify the images from the ImageNet dataset [78]. It contains 13 convolutional layers and 3 fully connected layers. We built an encoding model for each neuron from the activations of layer 14 (the first fully-connected layer). Layer 15 but not the output layer yielded similar predictive power. We did not modify or refit this CNN to predict neural responses. Instead, we ran nonlinear Poisson regression (XGBoost) to predict each neuron's response to an image from the values of layer 14 when the VGG network was given the same image. We found XGBoost to offer better predictions than other Poisson regression models. The final model thus takes a fixation image as input, runs the image through 14 layers of the VGG16 CNN, and then through a trained instance of XGBoost to predict the spike rate of a neuron. We call the combination of the CNN model and the trained XGBoost for each neuron the "CNN model".

The CNN model could then be used to build tuning curves. We conceptualized this as extracting the average first-order effect of hue upon the responses of this model to natural images. We perform the following cross-validated procedure for each of 8 bins of hues. First, we train the CNN model (i.e. train the XGBoost regressor) on the training set of the natural image dataset. We then modify the test set images by slightly desaturating all pixels whose hue lies within the current hue bin. The bins were chosen to be large (8 in instead of 16) to so as to be less affected by pixel noise and to speed computation. We desaturated by moving towards the L axis of the LUV color space, the same color space in which we define hue. For robustness, we modified images at each of many desaturation levels, ranging from 5% to 100% desaturation. We then obtained the predictions of the CNN model to the original test set and also for each modified, desaturated test set, and take the average difference of these two predictions across all images. This process is repeated in an 8-fold cross-validation procedure, so that each image serves as the test set once. The resulting series of average differences can be plotted against the desaturation. The slope of this line represents the average first-order contribution of that bin of hues to the images in the dataset. Note that the value of slope reflects the scale the x-axis, which represents the parameterization of the desaturation percentage. It is best to think of the units of slope as arbitrary; the important result is the relative value of the slope between hues. Finally, the process was repeated for each bin of hues, resulting in the tuning curve to hue.

We sought to validate this procedure on simulated data. One important aspect is that predictions are made on images that are as close to the distribution of images in the training set as possible. Since images in which a single bin of hues are desaturated by 5% are visually indistinguishable from the originals, this is not likely to be a concern. Nevertheless, we observed whether this method would be able to reconstruct the hue tuning of simulated neurons. We constructed 20 simulated neurons that responded linearly to the hues present in a receptive field. Each neuron was cosine tuned with a randomly selected hue angle. Linear regression could perfectly reconstruct the hue tuning of these simulated neurons, as expected. The CNN method could also reconstruct the tuning curves, though less well than linear regression (as indicated by the spread of cross-validated pseudo-R^2 values, Supp. Fig. 3). If linear tuning curves do exist, then, the CNN method would be able to reconstruct them.

Calculation of error bounds

Each estimate of a tuning curve represents, in essence, a summary statistic of noisy data. To estimate error bounds on tuning curves, we relied on the nonparametric method of bootstrapping across trials, or for summary statistics of the

entire neural population, additionally bootstrapping across neurons. Since the uniform field hue tuning curves used for correlations were simple averages of spike rates, binned over hue, we bootstrapped across trials to compute the confidence intervals. The natural scene tuning curves for the GLM and nonlinear methods represented the predicted response to single hues. For these methods, we computed uncertainty bounds on their predictions to single hues by retraining the methods on resampled datasets (with replacement) and selecting the 5th and 95th percentiles of the predicted output for each bin. For the CNN method, the tuning curves were calculated from linear fits of the difference in test set predictions as a function of hue bin desaturation. The difference in predictions was noisy across images, with large changes predicted for some images but small changes predicted for other images. This noise presented as uncertainty in the linear fit to the data. The error on the CNN tuning curve, then, represented the uncertainty in the linear fit to the test set predictions.

The uncertainty on each of the tuning curves was then propagated into the correlation between the natural scene and uniform field tuning curves. This was again done through bootstrapping. For a given natural scene/uniform field correlation, we correlated the natural scene and uniform field tuning curves from hundreds of model fits upon resampled data, yielding a large distribution of correlations. We then reported the mean, 5th, and 95th percentiles of this distribution. The uncertainty of the mean across neurons included a bootstrap across the trials used to build the tuning curves for each neuron, followed by a bootstrap across neurons.

Acknowledgements

We are grateful to Samantha Schmitt for assistance with data collection and spike sorting. K.K., P.R., H.F., and A.B. acknowledge support from NIH grants MH103910 and EY021579. M.S. was supported by NIH EY022928, MH118929, EB026593, and NSF NCS 1734901.

Author Contributions

A.B. prepared the modeling methodology, conducted analyses, prepared figures, and wrote the initial manuscript. P.R. and H.F. conceptualized the research, designed the experimental protocol, curated the data, prepared methodology

and initial analyses, and edited the manuscript. M.S. provided funding and resources for the experiment, designed the experimental protocol, performed the data collection, advised the analysis methodology, and edited the manuscript. K.K. conceptualized the project, provided funding, advised the experimental protocol, supervised the data analysis, and edited the manuscript.

Competing Interests

The authors declare no competing interests.

Data Availability

All electrophysiological data is available upon request. Data can be provided in various stages of preprocessing to expedite replication: as spike times in each session, or as spike counts paired with raw images, with fixation-centered images, or with extracted hue histograms.

Code Availability

Our in-house code for spike sorting is available at https://github.com/smithlabvision/spikesort, and the code used to process sorted data, run analyses, and create figures is available at https://github.com/KordingLab/V4py. Our analyses depended on Python v2.7 and a large number of open-source projects that can be found listed in https://github.com/KordingLab/V4py/requirements.txt, including Pandas v0.23, Tensorflow v1.1, Numpy v1.11, and Keras v2.2.

References

1. Carandini, M., Demb, J.B., Mante, V., Tolhurst, D.J., Dan, Y., Olshausen, B.A., Gallant, J.L., and Rust, N.C. (2005). Do we know what the early visual system does? Journal of Neuroscience *25*, 10577-10597.
2. Hegdé, J., and Van Essen, D.C. (2003). Strategies of shape representation in macaque visual area V2. Visual neuroscience *20*, 313-328.
3. Pasupathy, A., and Connor, C.E. (2001). Shape representation in area V4: position-specific tuning for boundary conformation. Journal of neurophysiology *86*, 2505-2519.
4. Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science *310*, 863-866.
5. Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences *111*, 8619-8624.
6. Connor, C.E., Brincat, S.L., and Pasupathy, A. (2007). Transformation of shape information in the ventral pathway. Current opinion in neurobiology *17*, 140-147.
7. Logothetis, N.K., and Sheinberg, D.L. (1996). Visual object recognition. Annual review of neuroscience *19*, 577-621.

8. Simoncelli, E.P., and Olshausen, B.A. (2001). Natural image statistics and neural representation. Annual review of neuroscience *24*, 1193-1216.
9. Vinje, W.E., and Gallant, J.L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. Science *287*, 1273-1276.
10. David, S.V., Vinje, W.E., and Gallant, J.L. (2004). Natural stimulus statistics alter the receptive field structure of v1 neurons. Journal of Neuroscience *24*, 6991-7006.
11. Touryan, J., Felsen, G., and Dan, Y. (2005). Spatial structure of complex cell receptive fields measured with natural images. Neuron *45*, 781-791.
12. David, S.V., and Gallant, J.L. (2005). Predicting neuronal responses during natural vision. Network: Computation in Neural Systems *16*, 239-260.
13. Felsen, G., and Dan, Y. (2005). A natural approach to studying vision. Nature neuroscience *8*, 1643.
14. Olshausen, B.A., and Field, D.J. (2006). What is the other 85 percent of V1 doing. L. van Hemmen, & T. Sejnowski (Eds.) *23*, 182-211.
15. Bashivan, P., Kar, K., and DiCarlo, J.J. (2019). Neural population control via deep image synthesis. Science *364*, eaav9436.
16. Cowley, B., Williamson, R., Clemens, K., Smith, M., and Byron, M.Y. (2017). Adaptive stimulus selection for optimizing neural population responses. In Advances in neural information processing systems. pp. 1396-1406.
17. David, S.V., Hayden, B.Y., and Gallant, J.L. (2006). Spectral receptive field properties explain shape selectivity in area V4. Journal of neurophysiology *96*, 3492-3505.
18. Oleskiw, T.D., Pasupathy, A., and Bair, W. (2014). Spectral receptive fields do not explain tuning for boundary curvature in V4. Journal of neurophysiology *112*, 2114-2122.
19. Touryan, J., and Mazer, J.A. (2015). Linear and non-linear properties of feature selectivity in V4 neurons. Frontiers in systems neuroscience *9*, 82.
20. Zeki, S.M. (1973). Colour coding in rhesus monkey prestriate cortex. Brain research *53*, 422-427.
21. Desimone, R., and Schein, S.J. (1987). Visual properties of neurons in area V4 of the macaque: sensitivity to stimulus form. Journal of neurophysiology *57*, 835-868.
22. Carlson, E.T., Rasquinha, R.J., Zhang, K., and Connor, C.E. (2011). A sparse object coding scheme in area V4. Current Biology *21*, 288-293.
23. Popovkina, D.V., Bair, W., and Pasupathy, A. (2019). Modeling diverse responses to filled and outline shapes in macaque V4. Journal of neurophysiology *121*, 1059-1077.
24. Watanabe, M., Tanaka, H., Uka, T., and Fujita, I. (2002). Disparity-selective neurons in area V4 of macaque monkeys. Journal of Neurophysiology *87*, 1960-1973.
25. Hinkle, D.A., and Connor, C.E. (2001). Disparity tuning in macaque area V4. Neuroreport *12*, 365-369.
26. Hinkle, D.A., and Connor, C.E. (2002). Three-dimensional orientation tuning in macaque area V4. Nature neuroscience *5*, 665.
27. Mountcastle, V.B., Motter, B., Steinmetz, M., and Sestokas, A. (1987). Common and differential effects of attentive fixation on the excitability of parietal and prestriate (V4) cortical visual neurons in the macaque monkey. Journal of Neuroscience *7*, 2239-2255.
28. Roe, A.W., Chelazzi, L., Connor, C.E., Conway, B.R., Fujita, I., Gallant, J.L., Lu, H., and Vanduffel, W. (2012). Toward a unified theory of visual area V4. Neuron *74*, 12-29.
29. Conway, B.R., Moeller, S., and Tsao, D.Y. (2007). Specialized color modules in macaque extrastriate cortex. Neuron *56*, 560-573.
30. Tanigawa, H., Lu, H.D., and Roe, A.W. (2010). Functional organization for color and orientation in macaque V4. Nature neuroscience *13*, 1542-1548.
31. Conway, B.R., and Tsao, D.Y. (2009). Color-tuned neurons are spatially clustered according to color preference within alert macaque posterior inferior temporal cortex. Proceedings of the National Academy of Sciences *106*, 18034-18039.
32. Li, M., Liu, F., Juusola, M., and Tang, S. (2014). Perceptual color map in macaque visual area V4. Journal of Neuroscience *34*, 202-217.
33. Bohon, K.S., Hermann, K.L., Hansen, T., and Conway, B.R. (2016). Representation of perceptual color space in macaque posterior inferior temporal cortex (the V4 Complex). Eneuro *3*, ENEURO. 0039-0016.2016.
34. Kusunoki, M., Moutoussis, K., and Zeki, S. (2006). Effect of background colors on the tuning of color-selective cells in monkey area V4. Journal of Neurophysiology *95*, 3047-3059.
35. Portilla, J., and Simoncelli, E.P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. International journal of computer vision *40*, 49-70.
36. Okazawa, G., Tajima, S., and Komatsu, H. (2015). Image statistics underlying natural texture selectivity of neurons in macaque V4. Proceedings of the National Academy of Sciences *112*, E351-E360.
37. Rosenthal, I., Ratnasingam, S., Haile, T., Eastman, S., Fuller-Deets, J., and Conway, B.R. (2018). Color statistics of objects, and color tuning of object cortex in macaque monkey. Journal of vision *18*, 1-1.
38. Osorio, D., and Vorobyev, M. (2008). A review of the evolution of animal colour vision and visual communication signals. Vision research *48*, 2042-2051.
39. Seung, H.S., and Sompolinsky, H. (1993). Simple models for reading neuronal population codes. Proceedings of the National Academy of Sciences *90*, 10749-10753.

40. Wang, Z., Stocker, A.A., and Lee, D.D. (2013). Optimal neural population codes for high-dimensional stimulus variables. In Advances in neural information processing systems. pp. 297-305.
41. Brunel, N., and Nadal, J.-P. (1998). Mutual information, Fisher information, and population coding. Neural computation *10*, 1731-1757.
42. Finkelstein, A., Ulanovsky, N., Tsodyks, M., and Aljadeff, J. (2018). Optimal dynamic coding by mixed-dimensionality neurons in the head-direction system of bats. Nature communications *9*, 3590.
43. Chelazzi, L., Della Libera, C., Sani, I., and Santandrea, E. (2011). Neural basis of visual selective attention. Wiley Interdisciplinary Reviews: Cognitive Science *2*, 392-407.
44. Motter, B.C. (1994). Neural correlates of attentive selection for color or luminance in extrastriate area V4. Journal of Neuroscience *14*, 2178-2189.
45. Mirabella, G., Bertini, G., Samengo, I., Kilavik, B.E., Frilli, D., Della Libera, C., and Chelazzi, L. (2007). Neurons in area V4 of the macaque translate attended visual features into behaviorally relevant categories. Neuron *54*, 303-318.
46. Connor, C.E., Preddie, D.C., Gallant, J.L., and Van Essen, D.C. (1997). Spatial attention effects in macaque area V4. Journal of Neuroscience *17*, 3201-3214.
47. Connor, C.E., Gallant, J.L., Preddie, D.C., and Van Essen, D.C. (1996). Responses in area V4 depend on the spatial relationship between stimulus and attention. Journal of neurophysiology *75*, 1306-1308.
48. Gallant, J.L., Connor, C.E., Rakshit, S., Lewis, J.W., and Van Essen, D.C. (1996). Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. Journal of neurophysiology *76*, 2718-2739.
49. Bushnell, B.N., Harding, P.J., Kosai, Y., Bair, W., and Pasupathy, A. (2011). Equiluminance cells in visual cortical area V4. Journal of Neuroscience *31*, 12398-12412.
50. Bushnell, B.N., and Pasupathy, A. (2012). Shape encoding consistency across colors in primate V4. Journal of neurophysiology *108*, 1299-1308.
51. Zeki, S. (1980). The representation of colours in the cerebral cortex. Nature *284*, 412-418.
52. Ziemba, C.M., Freeman, J., Movshon, J.A., and Simoncelli, E.P. (2016). Selectivity and tolerance for visual texture in macaque V2. Proceedings of the National Academy of Sciences *113*, E3140-E3149.
53. Grill-Spector, K., and Weiner, K.S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. Nature Reviews Neuroscience *15*, 536.
54. DiCarlo, J.J., and Cox, D.D. (2007). Untangling invariant object recognition. Trends in cognitive sciences *11*, 333-341.
55. Heeger, D.J. (1992). Normalization of cell responses in cat striate cortex. Visual neuroscience *9*, 181-197.
56. Knierim, J.J., and Van Essen, D.C. (1992). Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. Journal of Neurophysiology *67*, 961-980.
57. Sillito, A., and Jones, H. (1996). Context-dependent interactions and visual processing in V1. Journal of Physiology-Paris *90*, 205-209.
58. Fitzpatrick, D. (2000). Seeing beyond the receptive field in primary visual cortex. Current opinion in neurobiology *10*, 438-443.
59. McClure Jr, J.P., and Polack, P.-O. (2019). Pure tones modulate the representation of orientation and direction in the primary visual cortex. Journal of neurophysiology *121*, 2202-2214.
60. Yeh, C.-I., Xing, D., Williams, P.E., and Shapley, R.M. (2009). Stimulus ensemble and cortical layer determine V1 spatial receptive fields. Proceedings of the National Academy of Sciences *106*, 14652-14657.
61. Heitman, A., Brackbill, N., Greschner, M., Sher, A., Litke, A.M., and Chichilnisky, E. (2016). Testing pseudo-linear models of responses to natural scenes in primate retina. bioRxiv, 045336.
62. McIntosh, L., Maheswaranathan, N., Nayebi, A., Ganguli, S., and Baccus, S. (2016). Deep learning models of the retinal response to natural scenes. In Advances in neural information processing systems. pp. 1369-1377.
63. Benjamin, A.S., Fernandes, H.L., Tomlinson, T., Ramkumar, P., VerSteeg, C., Chowdhury, R.H., Miller, L.E., and Kording, K.P. (2018). Modern machine learning as a benchmark for fitting neural responses. Frontiers in computational neuroscience *12*.
64. Ponce, C.R., Xiao, W., Schade, P., Hartmann, T.S., Kreiman, G., and Livingstone, M.S. (2019). Evolving super stimuli for real neurons using deep generative networks. bioRxiv, 516484.
65. Felsen, G., Touryan, J., and Dan, Y. (2005). Contextual modulation of orientation tuning contributes to efficient processing of natural stimuli. Network: Computation in Neural Systems *16*, 139-149.
66. Pospisil, D.A., Pasupathy, A., and Bair, W. (2018). 'Artiphysiology'reveals V4-like shape tuning in a deep network trained for image classification. Elife *7*, e38242.
67. Morcos, A.S., Barrett, D.G., Rabinowitz, N.C., and Botvinick, M. (2018). On the importance of single directions for generalization. arXiv preprint arXiv:1803.06959.
68. Smith, M.A., and Sommer, M.A. (2013). Spatial and temporal scales of neuronal correlation in visual area V4. Journal of Neuroscience *33*, 5422-5432.
69. Snyder, A.C., Morais, M.J., and Smith, M.A. (2016). Dynamics of excitatory and inhibitory networks are differentially altered by selective attention. Journal of neurophysiology *116*, 1807-1820.
70. Shoham, S., Fellows, M.R., and Normann, R.A. (2003). Robust, automatic spike sorting using mixtures of multivariate t-distributions. Journal of neuroscience methods *127*, 111-122.
71. Kelly, R.C., Smith, M.A., Samonds, J.M., Kohn, A., Bonds, A., Movshon, J.A., and Lee, T.S. (2007). Comparison of recordings from microelectrode arrays and single electrodes in the visual cortex. Journal of Neuroscience *27*, 261-264.

72. Brainard, D.H., and Vision, S. (1997). The psychophysics toolbox. Spatial vision *10*, 433-436.
73. Bondar, I.V., Leopold, D.A., Richmond, B.J., Victor, J.D., and Logothetis, N.K. (2009). Long-term stability of visual pattern selective responses of monkey temporal lobe neurons. PloS one *4*, e8222.
74. McMahon, D.B., Jones, A.P., Bondar, I.V., and Leopold, D.A. (2014). Face-selective neurons maintain consistent visual responses across months. Proceedings of the National Academy of Sciences *111*, 8251-8256.
75. Cameron, A.C., and Windmeijer, F.A. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. Journal of Econometrics *77*, 329-342.
76. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software *33*, 1.
77. Chen, T., and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. (ACM), pp. 785-794.
78. Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.