The DeepTune framework for modeling and characterizing neurons in visual cortex area V4

Reza Abbasi-Asl[a,1], **Yuansi Chen**[b,1], **Adam Bloniarz**[b], **Michael Oliver**[c], **Ben D.B. Willmore**[c], **Jack L. Gallant**[c,d], **and Bin Yu**[a,b,2]

[a]Department of Electrical Engineering and Computer Sciences; [b]Department of Statistics; [c]Helen Wills Neuroscience Institute; [d]Department of Psychology, University of California, Berkeley, California 94720, USA

This manuscript was compiled on October 10, 2018

Deep neural network models have recently been shown to be effective in predicting single neuron responses in primate visual cortex areas V4. Despite their high predictive accuracy, these models are generally difficult to interpret. This limits their applicability in characterizing V4 neuron function. Here, we propose the DeepTune framework as a way to elicit interpretations of deep neural network-based models of single neurons in area V4. V4 is a midtier visual cortical area in the ventral visual pathway. Its functional role is not yet well understood. Using a dataset of recordings of 71 V4 neurons stimulated with thousands of static natural images, we build an ensemble of 18 neural network-based models per neuron that accurately predict its response given a stimulus image. To interpret and visualize these models, we use a stability criterion to form optimal stimuli (DeepTune images) by pooling the 18 models together. These DeepTune images not only confirm previous findings on the presence of diverse shape and texture tuning in area V4, but also provide rich, concrete and naturalistic characterization of receptive fields of individual V4 neurons. The population analysis of DeepTune images for 71 neurons reveals how different types of curvature tuning are distributed in V4. In addition, it also suggests strong suppressive tuning for nearly half of the V4 neurons. Though we focus exclusively on the area V4, the DeepTune framework could be applied more generally to enhance the understanding of other visual cortex areas.

computational neuroscience | visual cortex | V4 | tuning | stability | convolutional neural network

Understanding the function of primate visual pathways is a major challenge in computational neuroscience. Along the ventral visual pathway, cortical area V4 is of particular interest. It is a large retinotopically-organized area located intermediate between the early primate visual cortex areas such as V1 and V2 and high-level areas in the inferior temporal (IT) lobe. V4 is believed to be crucial for visual object recognition and visual attention, but its functional role remains mysterious. Computational studies of primary visual cortex have produced powerful quantitative models of V1 (1). Contrastingly, area V4 is more difficult to model computationally than V1. This is mainly due to its highly nonlinear response (2) and diverse tuning properties (3).

To understand the tuning properties of V4 neurons, one dominant traditional approach is to use handcrafted and limited synthetic stimuli (e.g. (4, 5)) to probe V4 neurons. For example, by comparing V4 neuron responses to Cartesian gratings with those to polar and hyperbolic (non-Cartesian) gratings, Gallant et al. (4, 6) found that V4 neurons are most selective for non-Cartesian gratings containing multiple orientations. Through a parameterized set of contour stimuli varying in angularity, curvature, and orientation, Pasupathy and Connor (5, 7) discovered that V4 neurons are selective to curved contour features. While such studies have successfully quantified V4 neuron responses to synthetic shapes, the tuning properties of most V4 neurons cannot be fully explored through these limited sets of stimuli (3).

An alternative approach to designing synthetic stimuli is using a large collection of natural images directly as stimuli. This approach reduces the difficulty in stimuli design, but creates a huge challenge in modeling. Specifically, it has been found that previously proposed simple and shallow computational models of V4 neurons perform poorly on natural images (3, 8, 9). For instance, David et al. (8) introduced the spectral receptive field (SRF) model to account for second order nonlinear response properties. The SRF model enhances our understanding of V4 orientation tuning properties, but its average prediction performance for the V4 neurons studied is far from satisfying (3). More recently, advances in deep convolutional neural networks (CNNs) with multiple layers of linear and non-linear operations have led to more accurate predictive models for neurons in V4 and IT (10–12). While this deep, convolutional and non-linear architecture is the key to the high predictive performance, it also makes the models difficult to interpret. This limits their usefulness in advancing neuroscience. A natural question arises: can we use these complex and accurate models to infer tuning properties of V4 neurons?

Significance Statement

Understanding how primates process visual information and recognize objects in an image is a major problem in neuroscience. Along the visual pathway, the midtier cortical area V4 is of particular interest. Despite its importance in the hierarchical organization of visual processing, its function remains elusive. Accurate deep neural network-based predictive models are built for responses of V4 neurons to natural image stimuli. While interpreting these models is traditionally difficult, we introduce the DeepTune framework to equip these complex models with stable interpretation and visualization. The DeepTune images provide rich, concrete and naturalistic characterizations of V4 neurons that refine significantly findings of previous studies. They hold promise as better natural input stimuli for future closed-loop experiments.

Author contributions: R.A., Y.C., A.B., M.O., B.D.B.W., J.L.G. and B.Y. designed research; B.D.B.W. and J.L.G. collected data; R.A., Y.C., A.B. and B.Y. contributed new analytic tools; R.A., Y.C., A.B. and B.Y. analyzed data; and R.A., Y.C., A.B., J.L.G. and B.Y. wrote the paper.

The authors declare no conflict of interest.

[1] R.A. and Y.C. contributed equally to this work.

[2] To whom correspondence should be addressed. E-mail: binyu@berkeley.edu

In this paper, we propose the DeepTune framework as a tool to visualize and interpret predictive models of single neurons. In order to make the interpretations be less dependent on arbitrary neural network architecture choices, we build an ensemble of 18 CNN-based models per neuron instead of a single model. The models vary in architecture, but all have comparable high and state-of-the art prediction accuracies. Each model uses a CNN to extract features from an input image. The CNN is pre-trained to perform object classification on the ImageNet dataset (13). The extracted features are then used as predictors to train a regularized linear regression model with the neuron firing rate as the response. This approach of applying a pre-trained model to a new prediction task is known as transfer learning (14). For each neuron, we then generate DeepTune images that are obtained via gradient optimization of the fitted models. Aggregating the DeepTune generation process from 18 models via a stability criterion, we further introduce the consensus DeepTune images for each neuron. We show that interpreting the components of DeepTune images that are consistent across 18 models and the consensus one can help better characterize the tuning property of a neuron and gain robustness against modeling choices. Finally, we perform population analysis of all DeepTune images from 71 neurons to illustrate the curvature tuning diversity and suppressive tuning in V4.

Results

We have recorded firing rates of 71 well isolated neurons in V4 from two awake-behaving male macaques. These recordings were previously used to study the sparseness of neural codes in the area V4 (but without predictive models) (15). The stimuli consist of a random sample of circular patches of grayscale digital photographs from a commercial digital library (Corel). Uniformly random sampled images without replacement were then concatenated into long sequences so that each 16.7 ms frame contained a random image from the library. When presented to the macaques, all images were centered on the estimated classical receptive field (CRF, see *SI Data Collection* for CRF estimation procedure). The image size was adjusted to be two to four times the CRF diameter (Figure 1-C). The training data set for each neuron contains 8,000-24,000 natural images (4,000-12,000 distinct ones * 2). Spike counts were measured at 60Hz, resulting in two measurements per image. For the holdout test dataset, 600 images (300 distinct ones * 2) were shown for each neuron in a fixed order, distinct from the images shown for the training dataset. The sequence of test images was repeated; for each neuron, each image in the test dataset was shown 8-10 times. The resulting spike counts were averaged to provide a more precise estimate of the expected spike count. In addition, repeats also allowed for estimating the amount of variance in the neuron explainable by the stimulus image (16) (see *SI Data Collection* for details).

CNN-based models are highly predictive of V4 neuron responses on natural stimuli. We introduce a transfer learning framework (Figure 1) to build predictive models in two stages for our V4 stimulus-response data as just described. For a given layer of a pre-trained CNN and for each input stimuli, in the first stage (Figure 1-A), we extract intermediate outputs from that layer of CNN as features. In the second stage (Figure 1-B), these features serve as predictors in a regularized linear regression (such as Ridge (17) or LASSO (18)) with time-lagged spike rates as the responses. Specifically, for one stimulus image at time t denoted as $\mathbf{z}_t \in \mathbb{R}^{s \times s}$ ($s = 227$ in the AlexNet CNN model (19)), the given layer of CNN transforms this image into a flattened feature vector $\mathbf{x}_t \in \mathbb{R}^d$ ($d = 256 \times 13 \times 13$ in the AlexNet-Layer2 CNN model). This feature transform is denoted as function $h : \mathbb{R}^{s \times s} \mapsto \mathbb{R}^d$. Since the responses of V4 neurons to a sequence of images are sensitive to the recent history of images shown to the subject, we build the models with multiple time lags. More precisely, we regress y_t against the training image features from last k frames of video prior to and including time t, i.e. $\mathbf{z}_t, \ldots, \mathbf{z}_{t-k+1}$.

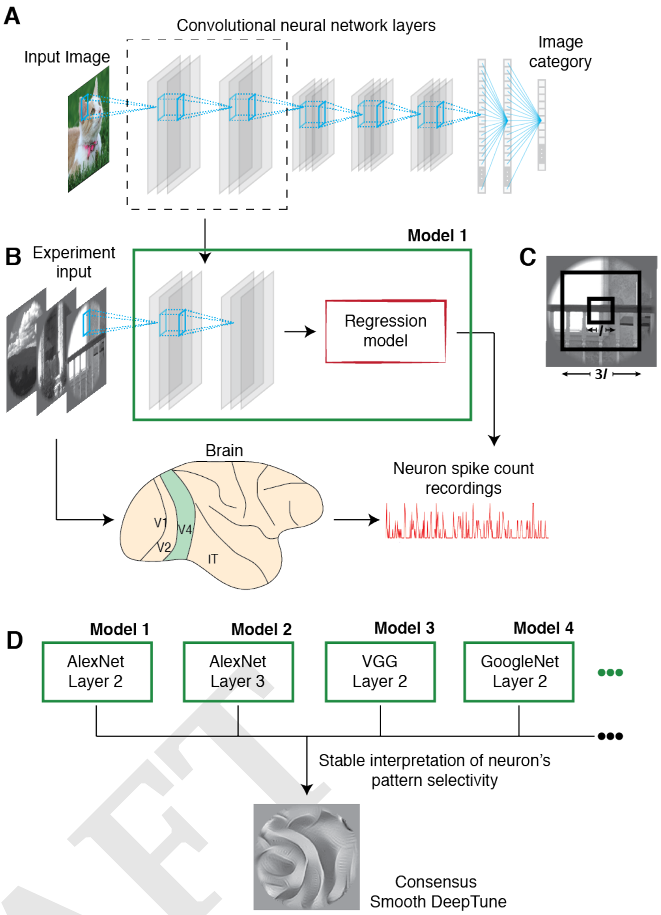

Fig. 1. DeepTune framework through transfer learning: first, we use features from pre-trained convolutional neural networks (CNNs) in regularized regression to predict (spike) firing rates of neurons in the visual area V4; second, stability-driven DeepTune images across 18 CNN-based predictive models are generated for interpretation. **A.** Architecture of a convolutional neural network (CNN) pre-trained to perform 1000-class image classification task on the ImageNet dataset (e.g. AlexNet). **B.** An input image is propagated forward in a fixed layer of the CNN, yielding a feature vector representation of the image. This vector is used to fit a regularized linear regression model to predict firing rates of each V4 neuron. **C.** The classical receptive field (CRF) during the experiment is set in the middle of the stimuli with width l while the whole image has the width $3l$. **D.** 18 accurate predictive models are obtained using features from layers 2, 3, 4 of three pre-trained AlexNet, GoogleNet, VGG, with either ℓ_1 (lasso) or ℓ_2 (ridge) regularized linear regression. DeepTune, a stability-driven interpretation and visualization framework of CNN-based model (across multiple such models) is proposed to characterize V4 neurons' tuning preferences (more details in the Results section 2). The consensus DeepTune image for one neuron (corresponds to Neuron 1 in Figure 3-A) is shown and displays a stable curvature pattern with edges forming an approximately ninety-degree angle.

The time lag k is set to be 9 (consisting frames at 0, 16.7, . . . , 133.6 ms) as in previous studies with similar data recordings (e.g. (8, 20)). Finally, our predictive model for a single neuron response takes the following form

$$F : \mathbb{R}^{s\times s\times k} \to \mathbb{R}$$

$$(\mathbf{z}_t, ..., \mathbf{z}_{t-k+1}) \mapsto \sum_{j=0}^{k-1} \boldsymbol{\beta}_{j+1}^T h(\mathbf{z}_{t-j}),$$

where $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) \in \mathbb{R}^{d\times k}$ are the regression parameters to be estimated and h is the fixed CNN feature transform. The model parameters are learned by solving the following regularized linear regression problem

$$\left(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_k\right) = \underset{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k}{\arg\min} \frac{1}{2}\sum_{t=k}^{T}\left(y_t - \sum_{j=0}^{k-1}\boldsymbol{\beta}_{j+1}^T h(\mathbf{z}_{t-j})\right)^2 +$$

$$\lambda_1 \sum_{j=1}^{k}\left\|\boldsymbol{\beta}_j\right\|_1 + \lambda_2 \sum_{j=1}^{k}\left\|\boldsymbol{\beta}_j\right\|_2^2.$$

If not specified in the rest of the paper, the regularization is taken to be ℓ_2 norm by setting $\lambda_1 = 0$ (Ridge). The analysis with ℓ_1 norm regularization (LASSO) by setting $\lambda_2 = 0$ to enforce sparsity is discussed in *SI Stability of Analysis*.

The CNNs used are pre-trained CNNs for classification tasks. They are trained based on a 1000-object classification task on the ImageNet dataset from the ImageNet Large Scale Visual Recognition Challenge (13). One legitimate concern of deploying neural networks in modeling is that interpretations about the models may depend on the details of the neural network architecture choices. To address this problem, we use three different neural network architectures to model V4 neurons: AlexNet (19), GoogleNet (21) and VGG (22). All three networks have high classification performance on ImageNet recognition challenge and are known to provide transferable image features in other computer vision tasks such detection and segmentation (14, 23). To vary the number of layers, we use features from layer two, three and four of each network. Later in this section, we show that using layer 1 and layers higher than layer 4 leads to lower prediction accuracies or has too large receptive fields not comparable with those of V4 neurons. Finally, on top of the CNN features, either Ridge or Lasso regression is used to predict the (spike) firing rates. As a result, we obtain 18 models for each neuron (3 nets × 3 layers × 2 regression models). Next we provide detailed prediction performance of these 18 models and compare them to previous models in the literature before we propose the stability-driven interpretation and visualization framework of DeepTune based on a stable aggregation of all 18 models.

To determine quantitatively how well our models describe the responses of each neuron, we test their performance on the holdout test set. All our models were estimated using the training data set. The correlation between the firing rates predicted by the model and the actual average firing rates on the test set is used as the prediction performance for all our 18 models. As a baseline for comparison, we also fit a V1-like Gabor wavelet model (24, 25). The Gabor wavelet model first extracts image features by applying a bank of linear Gabor wavelet filters to the input image at varying orientations, spatial frequencies and phases, followed by half-wave rectification and a compressive nonlinearity, then regresses the responses of each neuron using Ridge regression (17).

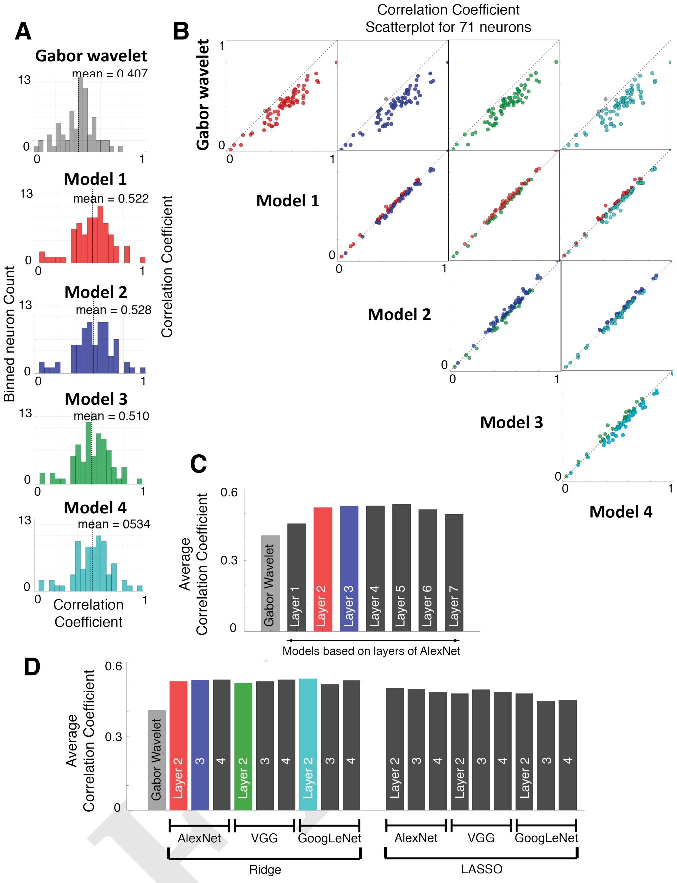

Fig. 2. CNN-based models outperform a V1-like Gabor wavelet model in terms of noise-corrected correlation coefficient (16) as the prediction performance measure. **A.** Histogram of noise-corrected correlation coefficients over the population of 71 V4 neurons for 4 models are shown, where the baseline model is a V1-like Gabor wavelet model, Model 1 corresponds to AlexNet-Layer2, Model 2 AlexNet-Layer3, Model 3 VGG-Layer2, and Model 4 GoogleNet-Layer2. Ridge regression is used in all 4 models. **B.** Scatter plots comparing noise-corrected correlation coefficients of 71 neurons between each pair among Models 1-4. Results for the other 14 models are shown in *SI Stability of Analysis, Figure S6*. **C.** Average prediction performance across 71 neurons for models from all 7 layers of AlexNet with ridge regression. The model based on AlexNet-Layer1 has the closest performance to that of the V1-like Gabor wavelet model; while models from layers 2 to 5 have higher predictive performance. **D.** Average prediction performance across 71 neurons for all 18 models. All 18 models perform similarly in prediction and much better than the Gabor wavelet model and the ridge-based models perform overall better than the lasso-based ones. Moreover, higher layers and more complex CNNs seem to result in worse performance for lasso, but not for ridge.

Our AlexNet-Layer2 (+Ridge) model has a average correlation coefficient of 0.44 (or 0.52 for noise-corrected correlation coefficient (16)) on the holdout test set. It achieves the state-of-the-art prediction accuracy for V4 neurons on natural image stimuli (8, 26). Comparing to (8), our average correlation coefficient is about 0.15 higher. As shown in Figure 2-D, all of the 18 models have average correlation coefficients higher than 0.42. For nearly all of the 71 V4 neurons, they are all more accurate than the V1-like Gabor wavelet model (with an average correlation coefficient 0.33). Due to space limitations, we plot the results only for 4 models, which are all based on AlexNet-Layer2, AlexNet-Layer3, VGG-Layer2, GoogleNet-Layer2 (and ridge) in Figure 2-A and 2-B. The first two models are chosen in order to demonstrate stability of prediction results and interpretations across different CNN layers, while the other two models are chosen to show stability across different CNN architectures (See *SI Stability of Analysis* for a complete comparison of the results from all 18 models). In Figure 2-C, we compare the average prediction performance for models from all 7 layers of AlexNet for 71 neurons. The model based on AlexNet-Layer1 has similar performance to that of the V1-like Gabor wavelet model; while models from layers 2 to 5 have much higher predictive performance (e.g. 0.44 for layer 2, 0.46 for layer 5). This justifies the recent finding (14) that the intermediate layers of pre-trained CNNs (on large-scale image classification tasks), like AlexNet, can extract more complex features than the first layer and Gabor wavelets.

In order to be consistent with the literature (8, 26, 27), we also report the proportion of explainable variance captured by a model. It attempts to control for differences in noise levels between experimental setups, individual neurons, and brain regions. We estimate the explainable variance through the noise-corrected correlation coefficient (16) using the repeated data in the holdout set (see *SI Stability of Analysis* for more information). Averaged over the 71 V4 neurons, the AlexNet-Layer2 and ridge model captures 30.3% of the explainable variance. This performance matches the 30% of computational models for area V2 (20). The unexplained portion of the response is very likely to have resulted from two factors: visual tuning properties not described by the AlexNet-Layer2 (and ridge) model and non-stimulus influences on the response. The latter is unlikely to be removed completely given our experimental setups (20). Note that the prediction task on the natural images in this paper is substantially harder than that on images with artificial objects overlaid in (10). Besides this work (10) on simpler natural image stimuli, our CNN-based models demonstrate a large improvement in prediction performance over previous works with natural image stimuli similar to ours (8, 26). In the next section, we take advantage of this high prediction accuracy to better characterize of V4 tuning properties via DeepTune images.

DeepTune as a naturalistic visual representation of tuning. It has long been challenging to fully characterize shape tuning properties in area V4. There are two main difficulties: the absence of highly predictive and biologically plausible computational models for the nonlinear response properties of V4 (3), and the lack of systematic methods to generate relevant complex natural stimuli to probe V4 neurons more efficiently. Given the state-of-the-art predictive performance of our CNN-based models, it is natural to ask whether these models could also provide a better characterization of shape tuning (e.g. angular, curvature or orientation tuning) or texture tuning in area V4. However, unlike existing studies using relatively simple Gabor wavelets (24, 25) or Fourier transform (8), complex nonlinear CNN features in our models make it extremely challenging to consistently interpret our models.

Inspired by computer vision advances in visualizing CNNs (28, 29), we introduce *DeepTune images* as a naturalistic visual representation of tuning for a V4 neuron. The DeepTune images are made of a collection of reconstructed images that jointly represent the shape tuning properties of a neuron. For each neuron and for each given model, a *DeepTune image* (or preferred DeepTune image) is obtained by optimizing over the input image space to maximize a regularized model output (predicted neuron response). Starting from a random image (e.g. white noise image with zero mean and fixed small variance), we use the gradient ascent method to gradually increase the model output until convergence. Formally, given a fixed predictive model at a particular time lag (the single lag time that causes best prediction performance in a 10% validation set split of the training set) $f : \mathbb{R}^{s\times s} \mapsto \mathbb{R}$, we seek an input image $\mathbf{z} \in \mathbb{R}^{s\times s}$ that minimizes the following objective function:

$$-f(\mathbf{z}) + \lambda_p \mathcal{R}_p(\mathbf{z}) + \lambda_{\mathrm{TV}} \mathcal{R}_{\mathrm{TV}}(\mathbf{z}).$$

The regularization terms are included to capture prior information about natural images. That is, the optimization search is constrained to be close to the set of smooth and naturalistic images (29). The specific regularization choices above are motivated by image denoising techniques (30) and by natural image statistics (31). The first regularizer $\mathcal{R}_p$ (the ℓ_p-norm of a vectorized image pixels) encourages the intensity of pixels to stay small. By choosing a large p ($p = 6$ in our analysis), this regularizer prevents the solution image from taking extremely large pixel values. The second regularizer $\mathcal{R}_{\mathrm{TV}}$ controls the total variation norm of an image. It encourages the image to be smooth and removes excessive high-frequency details (see *SI Methods* for more information).

The collection of DeepTune images is constructed from all 18 predictive models. In addition, we verify that 10 independent random initializations of starting images do not change the output much (see *SI Stability of Analysis*). Similarly, an inhibitory DeepTune is obtained by minimizing instead of maximizing the model output. We note that the DeepTune images differ from the traditional receptive fields in neurophysiology (24, 32) in two ways: multiple images are used to describe tuning properties of a single neuron; they are more naturalistic representations of tuning with a higher resolution.

Figure 3-A shows the DeepTune images from 4 of our 18 models built for Neuron 1. We visually observe that these DeepTune images share a stable curvature pattern with edges forming an angle of nearly 90 degrees. The rest 14 DeepTune images produced from the other 14 models differ slightly, but the main curvature pattern remains relatively stable (see *SI Stability of Analysis*). That is, the curvature angle stays close to 90 degrees and the spatial location of the curvature pattern remains at left side of the image. To further quantify the curvature angle and spatial frequency, we compare the power spectral densities (PSD) of these DeepTune images in Figure 3-B. All four DeepTune images share a strong and stable frequency component in the range of 45 to 135 degrees with spatial frequencies of 2 to 5 cycles per receptive field (green). Note that the high frequency components from the Model-4 DeepTune image are not consistent with the other three models. Especially, GoogleNet-Layer2 model has high frequency components that are not present in three other models. Therefore these components likely reflect noise and should be discounted. In Figure 3-C, we visualize the spectral receptive field (SRF) model (8) for Neuron 1. The SRF visualization shows the frequency components of the stimulus image selected by SRF model. The color map (red-blue) is chosen to be different from that of the DeepTune Fourier transform (green-pink). The color map difference serves a reminder of the difference between PSD and SRF. As observed from the DeepTune image PSD, the SRF model also shows that Neuron

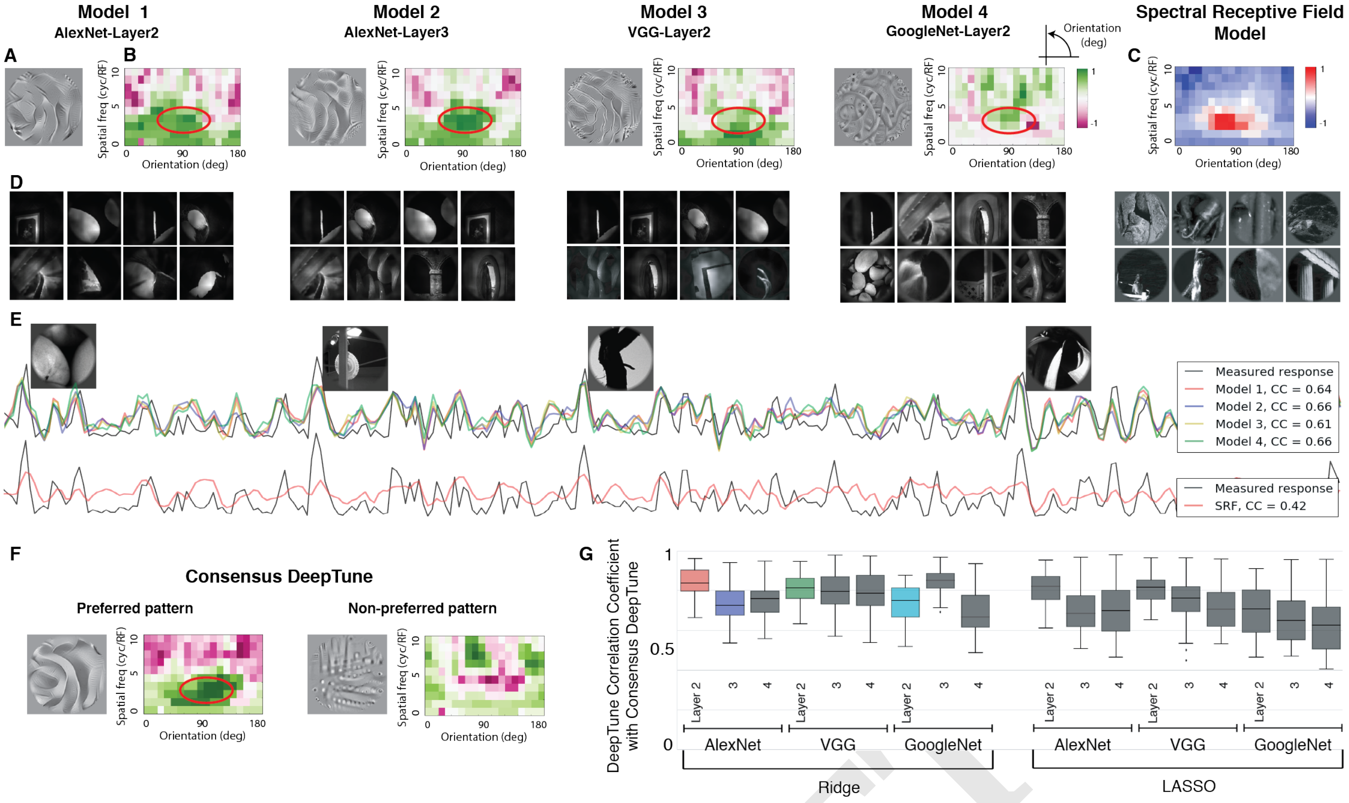

Fig. 3. DeepTune images from four of our 18 models built for Neuron 1. **A.** DeepTune images based on Models 1-4 for Neuron 1. These images share a visually stable curvature pattern with edges forming an approximately ninety-degree angle. **B.** Power spectral densities (PSDs) of the DeepTune images in polar coordinates. Through the PSDs, all four DeepTune images share a strong and stable frequency component in the range of 45 to 135 degrees with spatial frequency of 2 to 5 cycles per receptive field (the green color). **C.** Visualization of spectral receptive field (SRF) (8) model for Neuron 1. The SRF visualization emphasizes in red the frequency components of the stimulus image selected by the SRF model. The pattern selectivity according to SRF is consistent with the stable part of the PSDs of DeepTune images (highlighted in red circles). **D.** Images from training set with the highest responses for Neuron 1. Similar curvature patterns to the DeepTune visualization are visible in these images. **E.** The measured and predicted (spike) firing rates in the test set from Models 1-4 as well as the SRF model for Neuron 1. Images from the test set with the highest responses are visualized on top of the corresponding spike rate. Similar curvature patterns are visible in these images. Correlation coefficients between the measured and predicted firing rates are shown in the right panel. All four models outperform the SRF model. **F.** The consensus DeepTune image for Neuron 1. Both excitatory, inhibitory DeepTune images and the corresponding PSDs are shown. The excitatory pattern based on the consensus DeepTune exhibits the curvature contour that is similar to those from the four models in panel A. The inhibitory pattern visually consists of lines orthogonal to the preferred curvature contour, confirmed via PSD visualization on the right. **G.** Each box-plot corresponds to a CNN-based model among the 18 models and is based on 71 raw-pixel correlation coefficients. Each such coefficient corresponds to a neuron and is calculated between the consensus DeepTune image and a DeepTune image from that model and for that neuron. DeepTune images from AlexNet-Layer2 and GoogleNet-Layer 3 have the highest similarity on average to the consensus DeepTune image.

1 exhibits a strong preference to the frequency component in the range of 45 to 135 degrees with spatial frequency of 2 to 5 cycles per receptive field. In addition to DeepTune and SRF, this curvature tuning is further supported by the curvature patterns in the images from training and test sets with the highest responses for Neuron 1 (Figure 3-D and E). Figure 3-E illustrates the measured and predicted firing rates in test set from the 4 models as well as the predicted firing rates from the SRF model. For this Neuron 1, our 4 models have similar prediction accuracies (correlations on the holdout set between 0.61 to 0.64), while the SRF model has difficulty capturing the peak firing rates as seen in the lower plot of Figure 3-E, with a corresponding correlation of 0.42.

In addition to the visual comparison of 18 distinct DeepTune images generated from 18 models, we introduce consensus DeepTune to capture in a single image the stable patterns across 18 models. The consensus DeepTune image is obtained via a similar optimization scheme as in the original DeepTune optimization for a single model, but with an aggregation of gradient information from all 18 models. The aggregated gradient maintains the stable components in the gradients and discounts the unstable components (more details in *SI Methods*). Both excitatory and inhibitory consensus DeepTune images for Neuron 1 are shown in Figure 3-F. The excitatory consensus DeepTune (Figure 3-F) exhibits curvature contour patterns that visually matches all 4 models (Figure 3-A). The power spectral density (PSD) to the right of the consensus DeepTune image in Figure 3-F similarly matches the individual models. This PSD displays strong frequency components in the range of 45 to 135 degrees with spatial frequencies of 2 to 5 cycles per receptive field. On the other hand, the inhibitory consensus DeepTune consists of lines orthogonal to the curvature contour (see *SI Stability of Analysis* for comparison with inhibitory DeepTune images from all 18 models). Some blobs are also visible in the inhibitory consensus DeepTune image, suggesting that the response of Neuron 1 is attenuated by blob-like texture patterns. This is further supported by observing that the inhibitory PSD contains strong high frequency components on the top center.

The consensus DeepTune image captures the stable components of DeepTune images across our 18 models. It can be visually observed that the DeepTune images from a number of individual models are very similar to the Consensus DeepTune (see *SI Stability of Analysis, Figure S8*). To quantify this similarity, we compute the Pearson correlation coefficient between pixel values of the consensus DeepTune and those of each DeepTune image. Figure 3-G visualizes boxplots of these correlation coefficients. Each boxplot corresponds to one of the 18 models and shows the distribution of 71 correlation coefficients for all 71 neurons for this model. The median correlations for all of the models are considerably high. The highest median correlation is 0.83 which is achieved by AlexNet-Layer2 and GoogleNet-Layer3 with ridge regression. Models with lasso tend to have lower similarities to the consensus DeepTune. Due to space limitations, in the subsequent sections we present by default the consensus DeepTune image as a stable representation of a V4 neuron's tuning property. Although a single consensus DeepTune image seems to be sufficient, the stability analysis across 18 DeepTune images are necessary to determine the spatial locations of the stable parts. This is to ensure that we identify only the stable locations of the consensus DeepTune image to be interpreted.

Model-selected CNN features highlight receptive fields. The DeepTune images described in the previous section treated the CNN-based model as an end-to-end network. In this section, we show that analyzing the intermediate stages of a CNN-based model for a neuron can provide further information. The regression weights and the CNN features are of main interest. This analysis not only provides an independent and alternative interpretation of V4 neurons, but also allows us to compare our results to previously studied spatial receptive fields of V4 neurons.

Taking AlexNet-Layer2 model as an example, we examine its regression weights (see *SI Stability of Analysis, Figure S12* for visualization of weights from other models). Regression weights with large magnitudes indicate high sensitivity of the neuron to particular image features. The AlexNet-Layer2 features are of dimension 256 × 13 × 13. They consist of 256 different convolutional filters that are spatially located on a grid of size 13×13. The corresponding regression weights at one time lag is of the same dimension. We examine the regression weights by asking the following two questions: where on the image are the regression weights with the largest magnitudes? What kinds of convolutional filters contribute the most to the prediction performance?

To answer the first question, we define an *average regression weight map* as the sum-of-squares pooling of regression weights on the CNN features. It is defined across the different convolutional filters and the time lags at each location on the 13 × 13 spatial grid. Formally, for each neuron, let $\hat{\beta}_{mijk}$ be the regression weight for filter m at spatial location (i, j) and lag k. Then the average regression weight map $\Phi \in \mathbb{R}^{13\times13}$ is defined as follows:

$$\Phi_{ij} = \sum_{m=1}^{256} \sum_{k=1}^{k} \hat{\beta}_{mijk}^2.$$

Figure 4-A shows the average regression weight map from the AlexNet-Layer2 model for 4 neurons. On the 13 × 13 grid map, lighter pixel color indicates higher weight map value. Maps from other models share stable shape and location (see *SI Stability of Analysis, Figure S11* for a comparison across models). For each neuron, the average regression weight map presents an estimate for the spatial receptive field. Maps for V4 neurons exhibit diverse shapes. For example, the receptive fields for Neurons 1 and 2 have round shapes, while those for neurons 3 and 4 form straight or curved band shapes. These CNN-based spatial receptive fields provide an alternative to (34) for showing diversity in the size and shape of the receptive fields of V4 neurons. These regression weight maps are also indicative of the regions where DeepTune images across 18 models share stable patterns. Figure 4-B displays the DeepTune images from the AlexNet-Layer2 model for the 4 neurons, along with the consensus DeepTune images in Figure 4-C. The corresponding inhibitory DeepTune image and consensus inhibitory DeepTune image are shown in Figure 4-D and E respectively. Looking at the patterns of the DeepTune images, Neuron 1 is tuned to the curvature-contour shapes with edges forming an approximately ninety-degree angle. Neuron 2 is tuned to blob-like patterns and textures. Neuron 3 is selective to curvature patterns with a strong diagonal line preference and Neuron 4 is tuned to corner-like shapes with edges forming ninety-degree angles. The tuning patterns shown via DeepTune are consistent with receptive field shapes shown in regression weight maps.

The second question is: which types of convolutional filters contribute the most to the prediction performance? To address this question, we quantify the importance of each convolutional filter by ℓ_2 pooling of the regression weights for a convolutional filter across spatial locations. Formally, for each neuron, the filter importance I_m of m-th convolutional filter is defined as follows,

$$I_m = \sum_{i=1}^{13} \sum_{j=1}^{13} \sum_{k=1}^{k} \hat{\beta}_{mijk}^2,$$

where $\hat{\beta}_{mijk}$ is defined as before. This filter importance index provides an independent view of neuron shape tuning through the most and the least important filters. To interpret the filter importance, a visualization of each convolutional filter in CNN is required. To this end, we adopt the filter visualization technique introduced by (28). For each filter, we show the 9 top image patches from the ImageNet training set that have the highest filter responses (see *SI Methods* for more details). These 9 top image patches are representative of what this convolutional filter is computing (28, 33). Taking Neuron 1 as an example, Figure 4-F and G show the top and bottom two filters among 256 filters in AlexNet-Layer2 model ranked by the filter importance index, I_m.

For each neuron, we observe that the top two filters capture essential image components corresponding to the tuning patterns shown in the DeepTune images. These tuning patterns are long curvatures for Neuron 1, blob-like patterns for Neuron 2, diagonal lines for Neuron 3, and corner-like shapes for Neuron 4. Comparing to the DeepTune images (Figure 4-B-C-D-E), the most important and least important CNN-features (Figure 4-C-H-I) provide an alternative interpretation of the excitatory and inhibitory tuning property of V4 neurons, respectively. Figure 4 shows that these two views (I_m based and DeepTune) are visually consistent.

The wide variety of shape and texture tuning in V4. So far we have demonstrated that V4 neurons can be selective to both

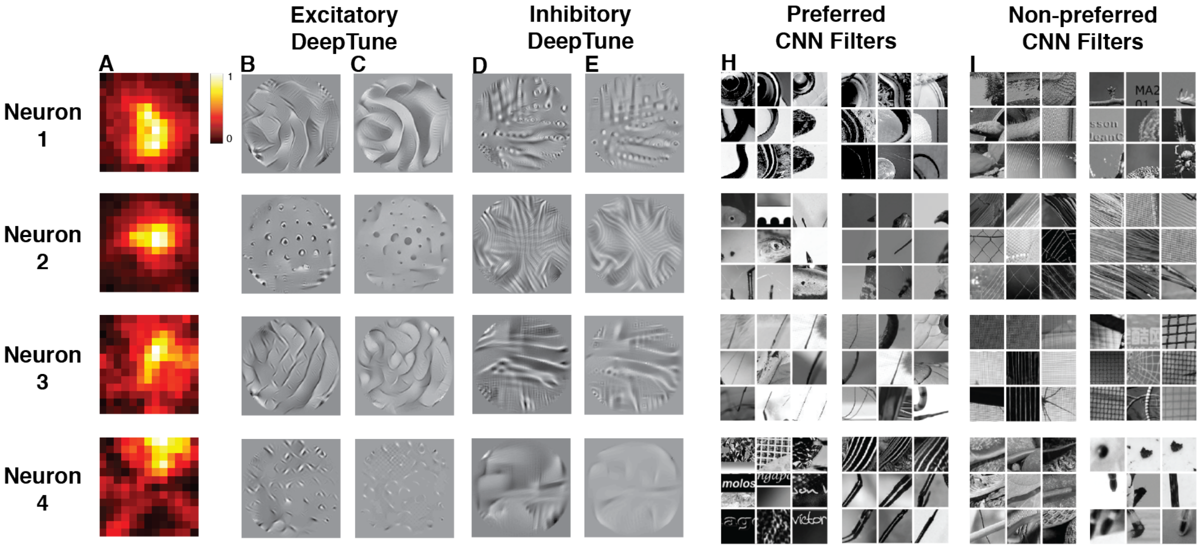

Fig. 4. For Neurons 1-4, a comparison of excitatory and inhibitory DeepTune images, average regression weight maps and selected CNN features. **A.** Average regression weight map based on the AlexNet-Layer2 model. For each neuron, the average regression weight map also exhibits stable patterns across models (see *Stability of Analysis*) and it highlights the receptive field of a neuron. **B.** Excitatory DeepTune images from the AlexNet-Layer2 Model. Neuron 1 is tuned to the curvature-contour shapes with edges forming an approximately ninety-degree angle. Neuron 2 is selective for blob-like patterns and textures. A DeepTune image for Neuron 3 shows selectivity to curvature patterns with a strong diagonal line preference. Neuron 4 is tuned to corner-like shapes with edges forming ninety-degree angles. The rest of the 17 models show consistent patterns as shown in other DeepTune images (see *SI Stability of Analysis, Figure S9*). **C.** Excitatory consensus DeepTune images based on all 18 models. **D.** Inhibitatory DeepTune images from the AlexNet-Layer2 Model. **E.** Inhibitory consensus DeepTune images based on all 18 models. **H.** Top two excitatory CNN filters based on the filter importance index. To visualize a convolutional filter from a CNN, the 9 top image patches are presented from the ImageNet training set that have the highest filter responses. These 9 top image patches are representative of what this convolutional filter is computing (28, 33). The top two selected CNN filters support the findings based on DeepTune images. For example, Neuron 1 is tuned for curved-contour patterns according to DeepTune images and its top CNN filters are those that activate on curvatures of similar shapes. Neuron 2 is selective for blob patterns and the top CNN filters activate respectively on blob pattern or pieces of a blob pattern. **I.** Top two inhibitory CNN filters based on the filter importance index.

shapes (e.g. contour or curvature patterns) and textures. The finding that V4 are tuned to both shapes and textures agrees with previous studies using synthetic stimuli: on the one hand, V4 neurons are shown to be tuned to orientation and spatial frequency of edges and linear sinusoidal gratings (35), non-Cartesian gratings (4, 6) and curvature of contours (5, 7); on the other hand, V4 is found to play a major role in processing textural information (36–38). In order to further understand area V4 as a population of neurons, we use DeepTune as a new tool to investigate proportions of V4 neurons that are tuned to shapes, to textures, and to other patterns of stimuli.

Based on visual inspections of their consensus DeepTune images, we manually clustered our 71 neurons into five categories: two texture categories (blob-like and corner-like patterns), one for curved contours, one for lines and a final category for complex patterns. Figure 5-A is the count histogram of these five categories and Figure 5-B displays DeepTune images for three example neurons in each category with high correlation coefficients (> 0.4). This visualization again confirms that both texture-tuned and contour-tuned neurons are present in area V4. In fact, among the 71 neurons considered in this study, about 40% of them are selective to textures and 30% of them prefer contour shapes. A finer manual categorization shows that among the ones selective to textures, half are tuned to blob-like patterns and the other half prefer corner-like patterns. Contour-selective neurons show preferences to either curvatures or straight lines like some typical V1 neurons but with larger receptive fields. The number of neurons selective to curved contours is twice of that selective to straight lines. We have also included in the last category the neurons tuned for complex patterns that are hard to describe in language and do not fall into previous categories. By displaying neuron tuning in a concrete and naturalistic manner, the DeepTune images extends the results in previous studies on V4 neuron selectivities (6, 39).

V4 curvature tuning to a full range of separation angles. It is suggested by Roe et al. (3) that diverse curvature tuning in V4 provides an efficient way to encode shapes. However, it is not yet clear that how different types of curvature tunings are distributed in the V4 population. Previously, artificial curvature stimuli have been used to probe the different angle tuning properties in area V4 (5, 7). These stimuli are constructed by joining two oriented line segments in a sharp corner or curve. These studies highlight the presence of bimodal orientation tuning with various separation angles. The preferred separation angle is defined in (5, 8) as the angle between the two most preferred oriented line segments passing through the center. The SRF analysis (8) also confirm bimodal orientation tuning in V4 by showing the presence of neurons tuned sharp corners. As for the distribution of different angles, Carson et al. (40) observed that not all curvatures are equally represented. They use sparse modeling of object coding to show that the strong representation of acute curvatures across the neural population. In this section, we investigate whether DeepTune images can concretize previous discoveries and provide visualization of V4 neurons tuned to different separation angles.

By visually inspecting the consensus DeepTune images of 71 V4 neurons, we first identified the 38 neurons that are tuned to curved contours, corner-like shapes and lines. Then we manually clustered these 38 neurons into four categories based

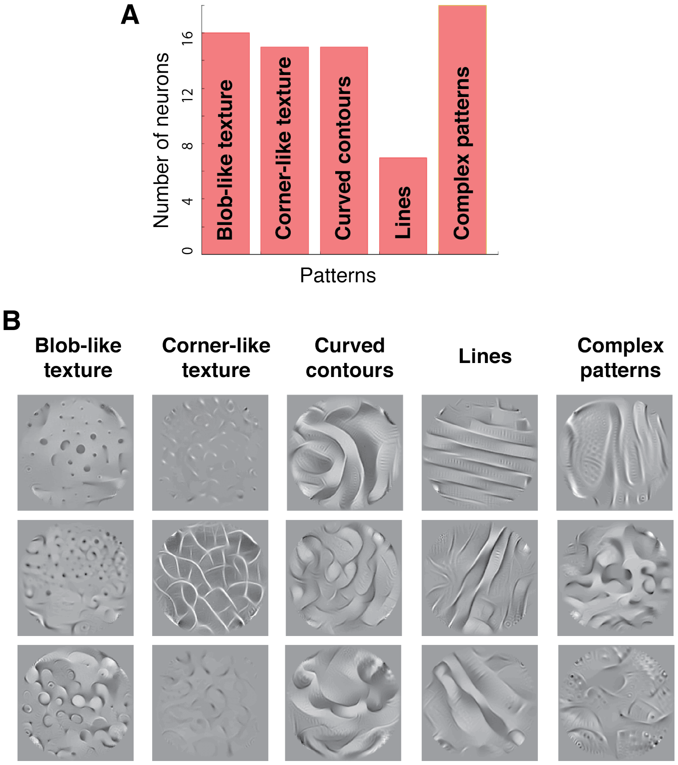

Fig. 5. Diversity of tuning among 71 V4 neurons. **A.** Neurons are manually categorized into five categories based on their DeepTune images. More than 40% of the neurons are selective to texture, half of which prefer blob-like textures and the other half prefer corner-like textures. About 30% of the neurons exhibit contour patterns, both curvature and straight lines. Neurons selective to curvatures are twice as the ones selective to straight lines. The rest of the neurons have selectivities to visually complex patterns. **B.** Examples of consensus DeepTune images for three neurons from each of the five categories.

on their separation angles of their curves (45°, 90°, 135°, 180°). Figure 6-A shows a count histogram of (excitatory) separation angle of the 71 V4 neurons. We observe that there is a strong presence of neurons with curvature tuning at less or equal to 90 $\circ$ separation angles (18 out of 71 neurons). Another 15 neurons are selective to blob-like textures that does not correspond to any particular angle. There are 18 neurons that are not selective to any clear angle or blob-like patterns.

To further support the separation angles for V4 neurons identified by looking at DeepTune images, we perform spectral receptive field (SRF) analysis (8) on our data and compare the angles identified by both analyses. In Figure 6-B and C, for each neuron, we display in one column the consensus DeepTune image and the SRF plot as in David et al. (8). The horizontal axes of the SRF show the orientation tuning of each neuron, with preferred component in red. In the SRF plot, according to (8), the separation angle corresponds to the difference between the top two orientation tuning peaks. We observe that the separation angle from the SRF plot are consistent with the ones from the DeepTune images. For example, for the bottom left neuron, both DeepTune and SRF show two orientation tuning peaks at about 70° and 120°. To summarize, the diversity of excitatory curved-contour patterns in fact matches the previous neurophysiological observations in V4 (5, 34, 40). Furthermore, our DeepTune images offer a concrete visualization of the bimodal orientation tuning

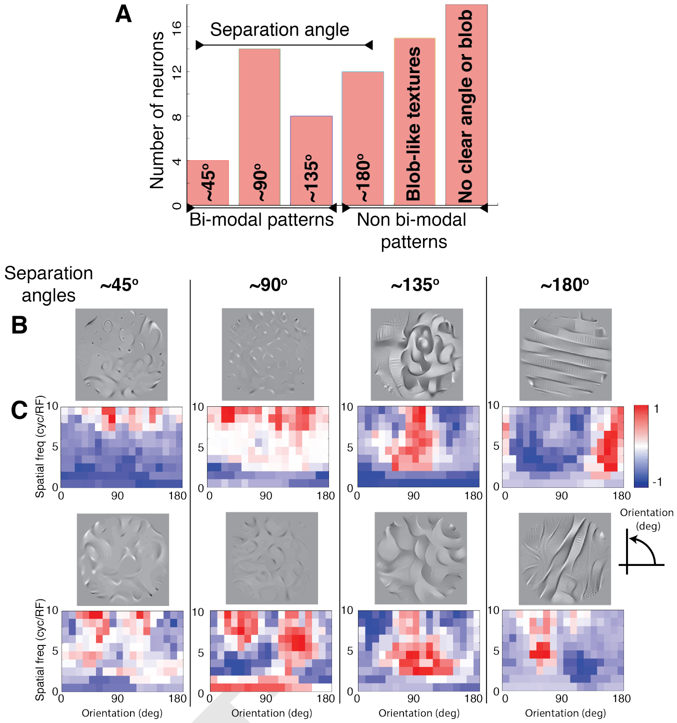

Fig. 6. Categorization of V4 neurons based on their separation angles. **A.** Neurons are manually categorized into six groups. The first four groups contain neurons tuned to patterns with separation angles of 45°, 90°, 135°, and 180°. These patterns are either contours or textures. About 20% out of 71 neurons are tuned to patterns with separation angles close to 90°. Another 20% of the neurons are selective to blob-like textures that do not correspond to any particular angle. The rest of neurons are not selective to any clear angle or blob-like patterns. **B.** The consensus DeepTune images for two example neurons in each of the first four categories. **C.** The corresponding spectral receptive field (SRF) (David et al (8)) visualization. The orientation tuning obtained via SRFs and consensus DeepTune images are consistent. while SRF predicts a neuron has tuning for a particular angle through Fourier analysis, the consensus DeepTune images offer concrete and detailed visualization of these tunings. For example, for the bottom left neuron, both our method and SRF show an orientation tunings of about 70° and 120°.

properties of many V4 neurons, refining earlier analysis.

Suppressive tuning discovery via inhibitory DeepTune. It is well known that V4 neurons have surround suppressive mechanisms (35, 41, 42) just like many other visual cortical areas (32, 43). Besides, recent study by Willmore et al. (20) found evidences for the presence of strong suppressive tuning to specific features in about half of the neurons in area V2. In addition, they show that this type of suppressive tuning is not caused merely by surround suppression and is not present in area V1. In this section, we investigate whether such strong suppressive tuning is also present in area V4.

To study the suppressive tuning in the area V4, we fit the Berkeley Wavelet Transform (BWT) model (20) to our data. The BWT-based model provides a nonlinear spatio-temporal receptive field (STRF) for each neuron (more details in *SI Methods*). We adopt the excitation index (EI) introduced in (20) as:

$$EI = \frac{\Sigma h^{+} - \Sigma h^{-}}{\Sigma h^{+} + \Sigma h^{-}}$$

where h^{+} and h^{-} are positive and negative weights respectively assigned to the wavelets in each STRF.

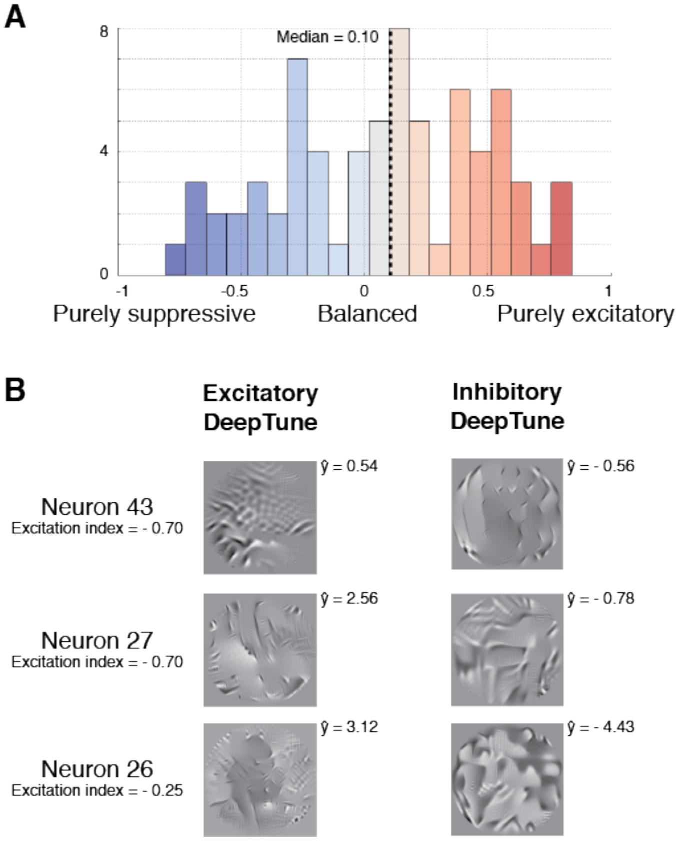

Fig. 7. Neurons in the primate cortical area V4 exhibit suppressive tuning. **A.** Histogram of BWT excitation index for 71 V4 neurons. 41% of the neurons show strong suppressive tuning. The median of excitation index for V4 neurons is 0.10. **B.** The excitatory and inhibitory DeepTune images for three neurons identified as suppressive by the BWT model. The neuron excitation index and response of the model to each DeepTune image is illustrated in the same panel. The neurons with suppressive tuning have much clearer suppressive DeepTune images than those without. $\hat{y}$ is the predicted model response obtained by feeding the DeepTune image through AlexNet-Layer2 model.

The BWT-based model has an average prediction correlation coefficient 0.33 for the 71 V4 neurons in the holdout test set. It is about 0.09 lower than the worst among 18 CNN-based models. While this model does not fully explain the non-linear property of V4 neurons, its accuracy is comparable to that of the same BWT model for V2 neurons (average correlation coefficient of 0.30) (20). Figure 7-A shows the histogram of excitation index for 71 V4 neurons. 41% of the neurons in V4 show suppressive tuning. The median of the excitation index for V4 neurons is 0.10. While the portion of neurons with suppressive tuning is 9% lower compared to that in V2, it is 29% higher than that in cortical area V1 (20).

Figure 7-B presents the excitatory and inhibitory consensus DeepTune images for three neurons identified as suppressive neurons according to the BWT model (on the left side of the histogram). The corresponding excitation indexes are shown below the neuron names. Recall that the excitatory DeepTune images are obtained via maximizing the model response (with appropriate regularization), while the inhibitory ones are obtained via minimizing the model response (with appropriate regularization). The neuron excitation index and response of the model to each DeepTune image are shown in the same panel. For example, $\hat{y} = 0.54$ means that the model predicted a firing rate of 0.54 spikes per sampling period (16.7ms). The DeepTune images provide a concrete visualization of the suppressive tuning in V4: The excitatory DeepTune images of these neurons are weak and/or blurry, while the inhibitory DeepTune images show sharper patterns. In the case of Neuron 43, while the excitatory DeepTune has blurry patterns, the inhibitory DeepTune exhibits a clear tuning to ninety-degree corner shapes in the right hand side of visual field. This means that a ninety-degree corner shape is likely to drive this neuron firing rates close to zero. Moreover, looking at the other inhibitory DeepTune images, both of neurons 27 and 26 have strong suppressive tuning to complex shapes with mid-range frequency.

Discussion

Prior work has demonstrated the power of deep CNNs in building accurate predictive models of neural responses in V4 (10, 11). In this work, we have similarly demonstrated that pre-trained CNNs give state-of-the-art results in modeling V4 neuron responses to natural images. We additionally have presented the DeepTune framework for eliciting stable visualizations and interpretations of these models. The generated visualizations are stable over modeling choices and randomness in the model training procedure.

Flexible visualization of optimal stimuli. The idea of computationally optimizing input stimulus to discover neuron tuning properties dates back to Carlson et al. (40). The evolutionary sampling method was used to optimize for the stimulus that causes the highest number of spikes. This work greatly expanded the search space of tuning patterns compared to previous methods that were based on handcrafted stimuli (4, 7). However, the evolutionary sampling method in (40) is constrained on limited concatenated Bezier splines. It can generate spline-based contours easily, but has difficulty for generating fine-scale texture stimuli. Our DeepTune images are generated from a regularized optimization directly over the input pixel values, and hence have an even larger search space that allows for more complex and naturalistic tuning patterns.

The resulting DeepTune population analysis demonstrates that V4 neurons are tuned to a huge variety of shapes as well as textures in different orientations. It also reveals that the tuning properties of many V4 neurons cannot be explained by simple edge and corner patterns. We see in Figure 7, for example, that even the stable part of the DeepTune images is difficult to describe in such simple terms. This suggests that tuning in area V4 is much more complex than that of V1 and than what can be described by handcrafted grating stimuli. Studies based on synthetic stimuli (4, 6, 7) may lack the expressive power to represent shapes of many V4 neuron receptive fields. Predictive modeling approaches as SRF (8) may not be sufficient to capture the complex tuning properties either. It provides only summary statistics such as spatial frequency and orientation about the receptive fields.

Distinctions in curvature selection revealed by DeepTune. Examining the DeepTune images of Neuron 3 and 4 in Figure 4, we see that both neurons are tuned to curvatures with similar edge orientations (two edge directions with a separation angle of ninety degrees). However, they have very distinct shape tuning properties apart from the orientation tuning summary statistics. Neuron 3 prefers a curvature-contour pattern with a ninety-degree angle and long edges. Neuron 4 prefers a corner-like repeated texture. This agrees with the study by Nandy et al. (34). It is suggested that the curvature selection of V4 neurons could arise for two reasons: systematic variation in fine-scale orientation tuning across spatial locations (like Neuron 3), and local tuning heterogeneity (like Neuron 4). Note that this type of refined result would be difficult to obtain via methods based on global Fourier analysis such as spectral receptive field (SRF) (8, 26). The 2D Fourier transform is spatial translation-invariant, meaning it is difficult to distinguish between Neuron 3 and Neuron 4 via SRF analysis.

DeepTune for future neurophysiology experiments. The DeepTune images for each V4 neuron are concrete and naturalistic. They are visually very similar to many input image stimuli. In other words, the DeepTune images are ready to be fed back to neurons as stimuli for confirmation or refutation of their characterizations of tuning properties in a closed experimental loop. Consequently, DeepTune images hold the promise to speed up the efficiency of data collection in V4 and other brain areas.

Materials and Methods

Electrophysiology. Extracellular recordings were made from well isolated neurons in parafoveal areas V4 (71 neural sites) of three awake, behaving male rhesus macaques (Macaca mulatta). Surgical procedures were conducted under appropriate anesthesia using standard sterile techniques (44). Areas V4 were located by exterior cranial landmarks and/or direct visualization of the lunate sulcus, and location was confirmed by comparing receptive field properties and response latencies to those reported previously (45, 46). All procedures were done in accordance with National Institutes of Health guidelines. See *SI Data Collection* for additional details.

Software Packages. The regularized linear regression analysis is performed using the SPAMS package (47). The neural network feature extraction and the DeepTune framework are implemented using the Caffe package (48). The pre-trained neural network architectures are from the Model Zoo of the Caffe package. See *SI Methods* for additional details.

ACKNOWLEDGMENTS. This research is supported in part by ONR Grant N00014-16-2664 , NSF Grants DMS1613002 and IIS 1741340, and the Center for Science of Information (CSoI), a US NSF Science and Technology Center, under grant agreement CCF-0939370. We would like to thank Bruno Olshausen and Friedrich Sommer for fruitful discussions and Anwar Nunez-Elizalde for helpful comments on the presentation of the paper.

1. Carandini M, et al. (2005) Do we know what the early visual system does? *Journal of Neuroscience* 25(46):10577–10597.
2. Touryan J, Mazer JA (2015) Linear and non-linear properties of feature selectivity in V4 neurons. *Frontiers in systems neuroscience* 9.
3. Roe AW, et al. (2012) Toward a unified theory of visual area V4. *Neuron* 74(1):12–29.
4. Gallant JL, Braun J, Van Essen DC (1993) Selectivity for polar, hyperbolic, and cartesian gratings in macaque visual cortex. *Science* 259(5091):100–103.
5. Pasupathy A, Connor CE (2002) Population coding of shape in area v4. *Nature neuroscience* 5(12):1332.
6. Gallant JL, Connor CE, Rakshit S, Lewis JW, Van Essen DC (1996) Neural responses to polar, hyperbolic, and cartesian gratings in area v4 of the macaque monkey. *Journal of neurophysiology* 76(4):2718–2739.
7. Pasupathy A, Connor CE (1999) Responses to contour features in macaque area v4. *Journal of Neurophysiology* 82(5):2490–2502.
8. David SV, Hayden BY, Gallant JL (2006) Spectral receptive field properties explain shape selectivity in area v4. *Journal of neurophysiology* 96(6):3492–3505.
9. Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL (2009) Bayesian reconstruction of natural images from human brain activity. *Neuron* 63(6):902–915.
10. Yamins DLK, et al. (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* 111(23):8619–8624.
11. Cadieu CF, et al. (2014) Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput Biol* 10(12):e1003963.
12. Yamins DLK, DiCarlo JJ (2016) Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience* 19(3):356–365.
13. Russakovsky O, et al. (2015) ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115(3):211–252.
14. Sharif Razavian A, Azizpour H, Sullivan J, Carlsson S (2014) Cnn features off-the-shelf: an astounding baseline for recognition in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops.* pp. 806–813.
15. Willmore BD, Mazer JA, Gallant JL (2011) Sparse coding in striate and extrastriate visual cortex. *Journal of neurophysiology* 105(6):2907–2919.
16. Schoppe O, Harper NS, Willmore BD, King AJ, Schnupp JW (2016) Measuring the performance of neural models. *Frontiers in computational neuroscience* 10.
17. Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12(1):55–67.
18. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.
19. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks in *Advances in neural information processing systems.* pp. 1097–1105.
20. Willmore BDB, Prenger RJ, Gallant JL (2010) Neural representation of natural images in visual area V2. *The Journal of neuroscience* 30(6):2102–2114.
21. Szegedy C, et al. (2015) Going deeper with convolutions in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* pp. 1–9.
22. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
23. Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? in *Advances in neural information processing systems.* pp. 3320–3328.
24. Daugman JG (1980) Two-dimensional spectral analysis of cortical receptive field profiles. *Vision research* 20(10):847–856.
25. Jones JP, Palmer LA (1987) An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of neurophysiology* 58(6):1233–1258.
26. Wu MCK, David SV, Gallant JL (2006) Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.* 29:477–505.
27. Roddey JC, Girish B, Miller JP (2000) Assessing the performance of neural encoding models in the presence of noise. *Journal of computational neuroscience* 8(2):95–112.
28. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks in *Computer vision–ECCV 2014.* (Springer), pp. 818–833.
29. Mahendran A, Vedaldi A (2015) Understanding deep image representations by inverting them in *Proceedings of the IEEE conference on computer vision and pattern recognition.* pp. 5188–5196.
30. Rudin LI, Osher S, Fatemi E (1992) Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena* 60(1-4):259–268.
31. Simoncelli EP, Olshausen BA (2001) Natural image statistics and neural representation. *Annual review of neuroscience* 24(1):1193–1216.
32. Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology* 195(1):215–243.
33. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H (2015) Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.
34. Nandy AS, Sharpee TO, Reynolds JH, Mitchell JF (2013) The fine structure of shape tuning in area v4. *Neuron* 78(6):1102–1115.
35. Desimone R, Schein SJ (1987) Visual properties of neurons in area v4 of the macaque: sensitivity to stimulus form. *Journal of neurophysiology* 57(3):835–868.
36. Merigan WH (2000) Cortical area v4 is critical for certain texture discriminations, but this effect is not dependent on attention. *Visual neuroscience* 17(6):949–958.
37. Arcizet F, Jouffrais C, Girard P (2008) Natural textures classification in area v4 of the macaque monkey. *Experimental brain research* 189(1):109–120.
38. Okazawa G, Tajima S, Komatsu H (2015) Image statistics underlying natural texture selectivity of neurons in macaque v4. *Proceedings of the National Academy of Sciences* 112(4):E351–E360.
39. Kobatake E, Tanaka K (1994) Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of neurophysiology* 71(3):856–867.
40. Carlson ET, Rasquinha RJ, Zhang K, Connor CE (2011) A sparse object coding scheme in area v4. *Current Biology* 21(4):288–293.
41. Schein SJ, Desimone R (1990) Spectral properties of v4 neurons in the macaque. *Journal of Neuroscience* 10(10):3369–3389.
42. Kondo H, Komatsu H (2000) Suppression on neuronal responses by a metacontrast masking stimulus in monkey v4. *Neuroscience research* 36(1):27–33.
43. Allman J, Miezin F, McGuinness E (1985) Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annual review of neuroscience* 8(1):407–430.
44. Vinje WE, Gallant JL (2002) Natural stimulation of the nonclassical receptive field increases information transmission efficiency in v1. *Journal of Neuroscience* 22(7):2904–2915.
45. Gattass R, Sousa AP, Gross CG (1988) Visuotopic organization and extent of V3 and V4 of the macaque. *The Journal of neuroscience* 8(6):1831–1845.
46. Schmolesky MT, et al. (1998) Signal timing across the macaque visual system. *Journal of neurophysiology* 79(6):3272–3278.
47. Mairal J, et al. (2014) Spams: A sparse modeling software, v2. 3. *URL http://spams-devel. gforge. inria. fr/downloads. html.*
48. Jia Y, et al. (2014) Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093.*