Systems/Circuits

A Hierarchical Statistical Model of Natural Images Explains Tuning Properties in V2

Haruo Hosoya[1,2] **and Aapo Hyvärinen**[3]

[1]Computational Neuroscience Laboratories, ATR International, Keihanna, Kyoto 619-0288, Japan, [2]Japan Science and Technology Agency, Presto, Kawaguchi, Saitama 332-0012, Japan, and [3]Department of Computer Science and Helsinki Institute for Information Technology, University of Helsinki, 00560 Helsinki, Finland

Previous theoretical and experimental studies have demonstrated tight relationships between natural image statistics and neural representations in V1. In particular, receptive field properties similar to simple and complex cells have been shown to be inferable from sparse coding of natural images. However, whether such a relationship exists in higher areas has not been clarified. To address this question for V2, we trained a sparse coding model that took as input the output of a fixed V1-like model, which was in its turn fed a large variety of natural image patches as input. After the training, the model exhibited response properties that were qualitatively and quantitatively compatible with three major neurophysiological results on macaque V2, as follows: (1) homogeneous and heterogeneous integration of local orientations (Anzai et al., 2007); (2) a wide range of angle selectivities with biased sensitivities to one component orientation (Ito and Komatsu, 2004); and (3) exclusive length and width suppression (Schmid et al., 2014). The reproducibility was stable across variations in several model parameters. Further, a formal classification of the internal representations of the model units offered detailed interpretations of the experimental data, emphasizing that a novel type of model cell that could detect a combination of local orientations converging toward a single spatial point (potentially related to corner-like features) played an important role in reproducing tuning properties compatible with V2. These results are consistent with the idea that V2 uses a sparse code of natural images.

Key words: **angle selectivity; end stopping; learning; natural image statistics; orientation integration; sparse coding**

Significance Statement

Sparse coding theory has successfully explained a number of receptive field properties in V1; but how about in V2? This question has recently become important since a variety of properties distinct from V1 have been discovered in V2, and thus a more integrative understanding is called for. Our study shows that a hierarchical sparse coding model of natural images explains three major response properties known in the macaque V2. We further provide a detailed analysis revealing the roles of different kinds of model cells in explaining the V2-specific properties. Our results thus offer the first sparse coding account for receptive field properties in V2 that has extensive biological relevance.

Introduction

The visual cortex encodes external inputs of tremendously high dimensionality with only a limited amount of neural resources. In light of this constraint, the statistical structure in natural inputs is likely to be exploited for achieving efficient encoding (Barlow, 1961). Sparse coding theory offers a candidate for such an encoding strategy, in which a network is adapted so that inputs are represented by a small number of neural activities. Indeed, it was demonstrated that such a model trained with appropriate naturalistic inputs exhibited Gabor filter representations similar to those of V1 simple cells (Olshausen and Field, 1996), as well as other V1 properties related to color, stereopsis, and motion (van Hateren and van der Schaaf, 1998; Hoyer and Hyvärinen, 2000); and extended models also explained contrast normalization (Schwartz and Simoncelli, 2001) and complex cell properties (Hyvärinen and Hoyer, 2000, 2001; Karklin and Lewicki, 2009). On the experimental side, some evidence of visual coding adapted to natural stimuli has been observed in monkey and ferret V1 (Vinje and Gallant, 2000; Berkes et al., 2011). Thus,

Received Dec. 18, 2014; revised May 19, 2015; accepted June 11, 2015.

Author contributions: H.H. and A.H. designed research; H.H. performed research; H.H. contributed unpublished reagents/analytic tools; H.H. analyzed data; H.H. and A.H. wrote the paper.

H.H. was supported by the Japan Science and Technology Agency, Presto program in "Decoding and Controlling Brain Information"; and partly by the Ministry of Internal Affairs and Communications, program in "Novel and innovative R&D making use of brain structures." A.H. was supported by the Academy of Finland, Centre-of-Excellence in Inverse Problems Research.

The authors declare no competing financial interests.

Correspondence should be addressed to Haruo Hosoya, 2-2-2 Hikaridai, Keihanna Science City, Kyoto Japan, 619-0288. E-mail: hosoya@atr.jp.

DOI:10.1523/JNEUROSCI.5152-14.2015

 0270-6474/15/3510412-17$15.00/0

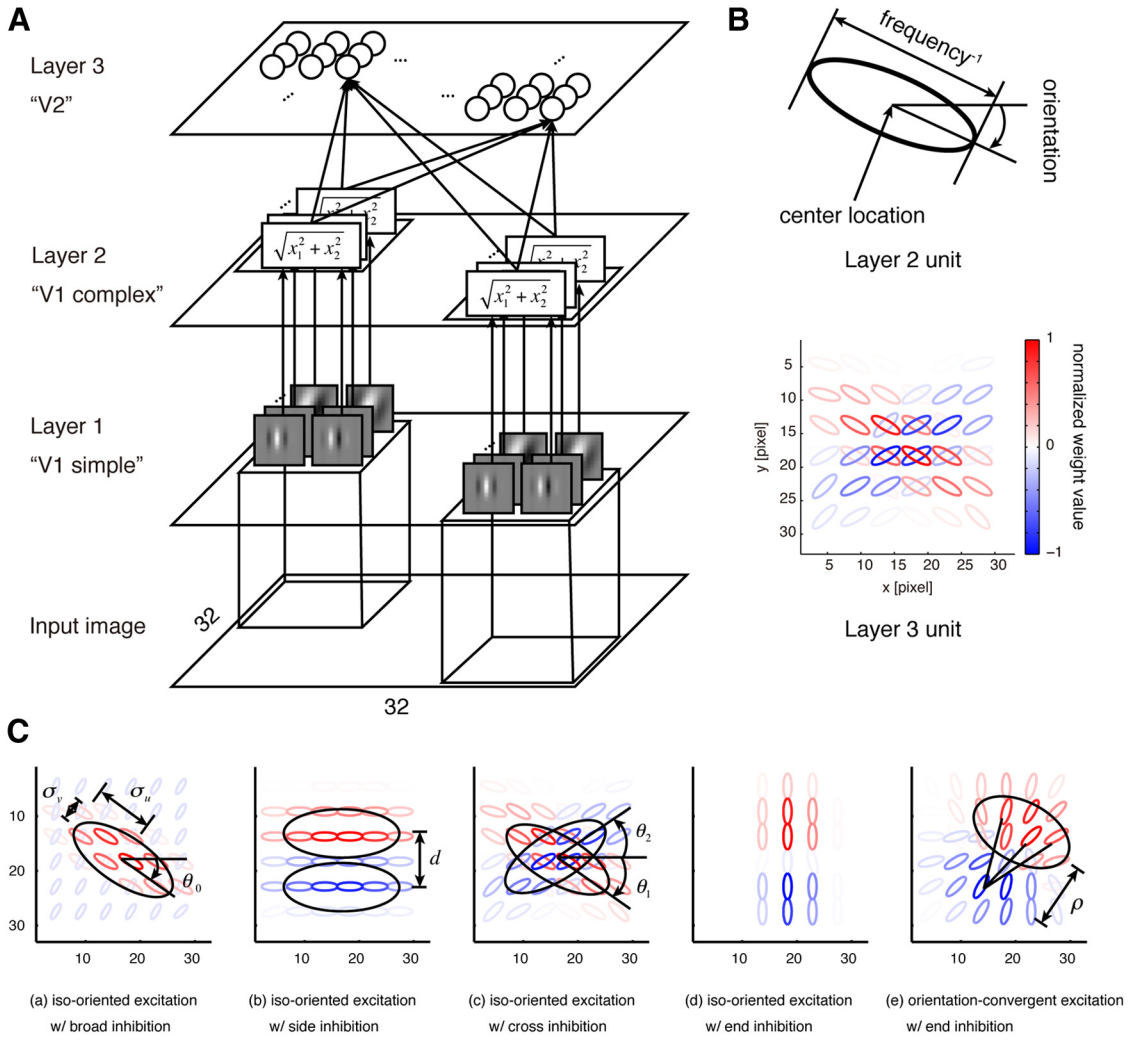

Figure 1. ***A***, A three-layered network architecture consisting of Layer 1 representing model V1 simple cells, Layer 2 representing model V1 complex cells, and Layer 3 representing model V2 cells (see Materials and Methods). ***B***, Visualization scheme. A Layer 2 unit is drawn as an ellipse with the orientation and the size proportional to the orientation and the inverse of frequency of the underlying Gabor filters, respectively. A Layer 3 unit is drawn as a set of ellipses corresponding to the Layer 2 units, with the colors indicating the normalized weight values (see the color bar). Only the maximum positive and minimum negative weights are shown at each position. ***C***, Illustration of five types of model units, overlaid with some parameters in the descriptive functions (see Materials and Methods).

representation in V1 seems tightly related to sparse coding of natural images. However, beyond V1, whether such a relationship exists or not is still unclear, despite a few preliminary studies (Lee et al., 2008; Hosoya, 2012; Gutmann and Hyvärinen, 2013).

Here, we present a theoretical investigation connecting sparse coding and neural representations in V2. The specific questions are twofold. First, if V2 is assumed to perform sparse coding of the output of V1, what are the emerging representations? Second, which tuning properties in V2 can such a model explain? These questions are important in particular because it is only relatively recently that several response properties distinct from V1 have been discovered in V2, and the precise nature of representations is much less well understood compared with V1. In such a situation, theoretical predictions would provide valuable insight into the actual neural representation.

To address the above questions, we trained a sparse coding model that took as input the output of a fixed V1-like model, which was in its turn fed a large variety of natural image patches as input. We then compared the response properties in the model and in actual macaque V2 with respect to the following three experimental protocols used in past macaque neurophysiological studies: (1) local orientation integration (Anzai et al., 2007); (2) angle selectivities (Ito and Komatsu, 2004); and (3) length and width suppression (Schmid et al., 2014). To gain further insight, we introduced a new analysis technique to classify the model V2 units according to their excitatory and inhibitory organization of local orientations and related these to the response properties.

We show that our model reproduced the aforementioned three major experimental results qualitatively, quantitatively, and robustly across various model variations. In addition, we provide detailed interpretations of the experimental data based on the modeling results, emphasizing the crucial role of a novel type of model cell exhibiting "orientation-convergent excitation with end inhibition" (potentially related to corner detection) in reproducing tuning properties compatible with V2.

Materials and Methods

Model architecture. We used a three-layer feedforward network model with the following architecture (Fig. 1*A*).

Layer 1 ("V1 simple") was a Gabor filter bank receiving a grayscale input image patch of size 32×32 pixels. The bank was pre-fixed and had all combinations of grid-arranged 6×6 center locations (at intervals of 4

pixels), 12 local orientations (at 15° intervals), three frequencies (0.25, 0.17, and 0.13 cycles/pixel, or, equivalently, 8.0, 5.3, and 4.0 cycles per 32 pixel patch width), and two phases (0° and 90°). Each Gabor filter with frequency f had Gaussian width and length both equal to $0.4/f$ (thus 1.6, 2.4, and 3.2 pixels for the three frequencies) and Euclidean norm equal to $f^{1.15}$ (thus, 0.20, 0.13, and 0.09 corresponding to the above three frequencies) in accordance with $1/f$ spectrum of natural images. The latter ensured the variances of the outputs of Layer 2 units (below) to be equal across frequencies. (Note that a Gabor pyramid could achieve a similar effect by its denser layout for higher-frequency units. However, we adopted the current design since this allowed for a precise adjustment of signal variance, which was crucial for appropriate learning. Also, visualization was clearer and formal analysis of learned representations was easier with the current design.) For computational efficiency, each filter was trimmed to its central 12×12 pixels (where the loss in norm was $<1.2\%$). Each Layer 1 unit computed, as output, the inner product of its associated Gabor filter and the input vector. This layer had a total of 2592 units.

Layer 2 ("V1 complex") formed a bank of standard energy models of phase-invariant V1 complex cells. The units in this layer were arranged in parallel to Layer 1, and had all combinations of the same 6×6 center locations, 12 local orientations, and three frequencies. Each unit first computed the Euclidian norm of the outputs of two Layer 1 units of the same center location, orientation, and frequency, but of different phases. Then, the mean of the outputs of all units was subtracted from each output. The last operation could be seen as a simplified implementation of contrast normalization. It did not qualitatively change the overall learned representations, but had an effect of reducing noise in the learned weights and adjusting the average of the weights to zero, which greatly simplified later formal classification. This layer had a total of 1296 units.

Layer 3 ("V2") performed sparse coding of Layer 2 outputs and had 800 units. Layers 2 and 3 were fully connected with weights of real value, and each Layer 3 unit computed the half-rectified value of the inner product of the associated weight vector and the output vector from Layer 2. The weight matrix W of size 800×1296 was subject to learning from natural image patches, as described below.

Learning method. For training Layer 3, we used ImageNet10K (Deng et al., 2010), a publicly available dataset of natural images containing ~9,000,000 images of 10,000 categories. Since our training method was unsupervised, the category labels associated with those images were unused. Each image was translated to grayscale, and resized so that a landscape image had a height of 128 pixels and a portrait image had a width of 128 pixels; the images with an original size smaller than 128×128 pixels were discarded. Then, each image was normalized to zero mean and unit variance.

During training, an image patch of size 32×32 pixels was repeatedly extracted at a random position from a randomly selected image. Low-contrast patches were discarded (variance, <0.32; acceptance rate, 97%) for numerical stability. Each extracted patch was again normalized to zero mean and unit variance. Then, the patch was fed to the network input layer, and the outputs of units in each layer were computed in a bottom-up manner.

The weight matrix W between Layers 2 and 3 was adapted to the statistics of Layer 2 outputs using a sparse coding principle. We used a particular sparse coding method based on independent component analysis (ICA; Hyvärinen et al., 2001) since ICA is known to be mathematically equivalent to the classic sparse coding model in certain conditions (Olshausen and Field, 1997), and since ICA generally offers estimation algorithms that are computationally efficient and fit well with a feedforward architecture. Among various ICA estimation methods, we adopted the score-matching method for overcomplete ICA (Hyvärinen, 2005), which can estimate more independent components than the input dimensionality, but with two important modifications. First, we chose to work on a huge number of image patches and therefore used a stochastic gradient method for score matching. Second, before applying ICA, we performed a strong dimension reduction on the Layer 2 outputs from 1296 to 100 dimensions (~13 times reduction) by principal component analysis (PCA). Although such an operation was rather unusual, the dimensions corresponding to low eigenvalues represented fine-grained structures in the input, and rejecting such dimensions had an effect of spatial pooling of subunits with similar orientation and frequency selectivities. Indeed, if a weaker dimension reduction was used, the pooling effect was diminished and the estimated features became smaller while their shapes remained similar. We chose to use 100 dimensions since this yielded the ratio of the average receptive field size of Layer 3 units to that of Layer 2 units comparable to physiology, whereas the ratio was much lower in the case of a larger number of dimensions (see Results). Here, the receptive field size was the diameter of the minimum circular region that covered all locations evoking 10% of the maximum response when presented with 3×3 pixel noise stimuli.

The details of the learning method were as follows. (1) The inputs (Layer 2 outputs) were centered and whitened with dimension reduction using a standard PCA algorithm. The PCA estimation was performed for 240,000 image patches. One hundred principal components corresponding to the largest eigenvalues were used for performing reduction of the original dimensionality of 1296. The variances of the remaining principal components were normalized to one, which amounts to whitening the dimension-reduced data. (2) A total of 800 independent components were estimated from the dimension-reduced, whitened inputs using a stochastic gradient algorithm for score matching (using log cosh nonlinearity) that worked over minibatches. Since 800 independent components were estimated from 100 input dimensions, the resulting representation was eight times overcomplete. The estimation was performed for 4,000,000 patches with a minibatch size of 500, and with an update rate started with 0.02 and halved after every 800,000 patches.

Putting these together, if we write E for the matrix of 100 (normalized) row eigenvectors (step 1), D for the diagonal matrix of the corresponding eigenvalues (step 1), and B for the matrix of 800 (normalized) row filter vectors estimated by overcomplete ICA on the dimension-reduced, whitened inputs (step 2), then the matrix W of row weight vectors between Layers 2 and 3 can be written as $W = BD^{-1/2}E$.

Visualization of the model units. After network training, properties of the learned model V2 (Layer 3) units were investigated by a series of analyses. The first was to visualize their internal representations. We consider here two formats of internal representations, a filter vector and a basis vector. The filter vector of ith unit is the ith row vector of the weight matrix W. The corresponding basis vector is the ith column vector of the basis matrix $A = E^T D^{1/2} B^T$. The latter matrix is essentially the inverse of the filter weight matrix W. (More precisely, the matrix A is an approximation of the pseudo-inverse $W^{\#} = E^T D^{1/2} B^{\#}$ of the generally nonsquare weight matrix W, assuming that the column vectors of B are nearly orthonormal, which was always the case for our results.) Note that the difference between the two matrices is precisely the exponent to D, which results in emphasizing coarser-grained structures in the basis vector (since principal components of Layer 2 outputs for higher eigenvalues are more amplified in the basis matrix A). This is why their structures are similar overall, albeit different in details (Fig. 2; see Fig. 5). Alternatively, A can be obtained by multiplying W^T with the covariance matrix (equal to $E^T DE$) from the left, which means that A can be seen as a smoothed version of W^T. This issue is rather classic in the theory of ICA, and a more detailed discussion can be found in the book by Hyvärinen et al. (2009).

The two formats are suitable for different purposes. On the one hand, the filter format provides a bottom-up view of the model, where the role of each unit is to measure the contribution of a certain local orientation pattern in a given natural image. On the other hand, the basis format gives a top-down view of the model, where the meaning of each unit is to generate local orientation features such that the sum of the features generated by all units constitutes a natural image. The filter format is more closely related to the details of the computation performed in the network, while the basis format gives a more intuitive interpretation in the space of natural image stimuli. Therefore, in the sequel, we work with the basis format whenever we informally explain and formally classify the model units; we often refer to the filter format when we describe the detailed response properties of the model units and compare them with experimental data.

To visualize a filter or basis vector, recall that each vector has 1296 dimensions, which correspond to the model V1 complex (Layer 2) units representing the combinations of 6×6 center locations, 12 local orien-

tations, and three frequencies. Thus, each vector is drawn as a set of ellipses, where the position, orientation, and size of each ellipse indicate the center location (in the visual field coordinates), orientation, and inverse of the frequency, respectively, of the corresponding model V1 complex unit; the color indicates the weight value normalized by the maximal absolute weight value (Fig. 1*B*). For readability, we show only the ellipses corresponding to the maximal positive (excitatory) weight and the minimal negative (inhibitory) weight at each location. Although this might appear to hide potentially important details, our visual inspection indicated that, for most units, the local weight pattern had only one positive peak and one negative peak at each position and frequency, and that the structure of orientation integration was very similar across frequencies. Therefore, our visualization method does not lose much information, while it allows us to display many units in a compact manner so that the tendency over the population can be exposed.

Unit classification. The learned model V2 units were classified according to the excitatory and inhibitory patterns of their basis vectors. In preliminary inspection, we identified the following five types: (1) iso-oriented excitation with broad inhibition; (2) iso-oriented excitation with side inhibition; (3) iso-oriented excitation with cross inhibition; (4) iso-oriented excitation with end inhibition; and (5) orientation-convergent excitation with end inhibition (Fig. 1*C*; see Results). To formally classify the model units according to these types, below we first define four descriptive functions corresponding to the four inhibition types (Fig. 1*C*). Here, we use Gaussian function $\phi(a; \mu, \sigma) = \exp\left[-\frac{(a-\mu)^2}{2\sigma^2}\right]$ with height 1, mean u, and standard deviation σ, and von Mises function $\psi(a; \mu, \sigma) = \exp\left[\frac{\cos 2(a-\mu)-1}{\sigma}\right]$ (cyclic in orientation, with period 180°) of height 1, mean u, and width σ. (Note that the free parameters in these functions and those defined below are written after a semicolon.)

Type I is iso-oriented excitation with broad inhibition (Fig. 1*C*,*a*), which is described as follows:

$$F_{\text{broad}}(x, y, \theta, f; x_0, y_0, \theta_0, f_0, \sigma_u, \sigma_v, \sigma_\theta, \sigma_f, A, b) = A \cdot \phi(u; 0, \sigma_u)\phi(v; 0, \sigma_v)\psi(\theta; \theta_0, \sigma_\theta)\phi(f; f_0, \sigma_f) + b$$

where the shifted and rotated position (u, v) is defined as $u = (x - x_0)\cos\theta_0 + (y - y_0)\sin\theta_0$ and $v = -(x - x_0)\sin\theta_0 + (y - y_0)\cos\theta_0$ (where u is along the envelope and v is orthogonal to it). The function F_{broad} thus expresses an excitatory subfield as a Gaussian envelope oriented at θ_0 with length σ_u and width σ_v around position (x_0, y_0). The envelope embeds, at each position, local orientations peaked at θ_0 with width σ_θ and frequencies peaked at f_0 with width σ_f. The baseline b, which is assumed to be a small negative value, describes a broad, nonspecific inhibitory pattern.

Type II is iso-oriented excitation with side inhibition (Fig. 1*C*,*b*), which is described as follows:

$$\begin{aligned} F_{\text{side}}(x, y, \theta, f; x_0, y_0, \theta_0, f_0, \sigma_u, \sigma_v, \sigma_\theta, \sigma_f, d, A) \\ = A \cdot \phi(u; 0, \sigma_u)\phi(v; 0, \sigma_v)\psi(\theta; \theta_0, \sigma_\theta)\phi(f; f_0, \sigma_f) \\ - A \cdot \phi(u; 0, \sigma_u)\phi(v; d, \sigma_v)\psi(\theta; \theta_0, \sigma_\theta)\phi(f; f_0, \sigma_f) \end{aligned}$$

where u and v are defined in the same way as Type I. The function F_{side} has an excitatory pattern that is similar to the function F_{broad} but has an additional inhibitory pattern with an identical shape to the excitatory pattern except that the center is shifted by distance d orthogonally to the envelope.

Type III is iso-oriented excitation with cross inhibition (Fig. 1*C*,*c*), which is described as follows:

$$\begin{aligned} F_{\text{cross}}(x, y, \theta, f; x_0, y_0, \theta_1, \theta_2, f_0, \sigma_u, \sigma_v, \sigma_\theta, \sigma_f, A) \\ = A \cdot \phi(u_1; 0, \sigma_u)\phi(v_1; 0, \sigma_v)\psi(\theta; \theta_1, \sigma_\theta)\phi(f; f_0, \sigma_f) \\ - A \cdot \phi(u_2; 0, \sigma_u)\phi(v_2; 0, \sigma_v)\psi(\theta; \theta_2, \sigma_\theta)\phi(f; f_0, \sigma_f) \end{aligned}$$

where $u_i = (x - x_0)\cos\theta_i + (y - y_0)\sin\theta_i$ and $v_i = -(x - x_0)\sin\theta_i + (y - y_0)\cos\theta_i$ for $i = 1,2$. The function F_{cross} has excitatory and inhibitory patterns that have identical shapes except for the orientations θ_1 and θ_2. (Note that the two orientations are not necessarily orthogonal.)

Type IV is iso-oriented or orientation-convergent excitation with end inhibition (Fig. 1*C*,*d*,*e*), which is described as follows:

$$\begin{aligned} F_{\text{end}}(x, y, \theta, f; x_0, y_0, \theta_0, f_0, \sigma_u, \sigma_v, \sigma_\theta, \sigma_f, d, \rho, A) \\ = A \cdot \phi(u; 0, \sigma_u)\phi(v; 0, \sigma_v)\psi(\theta; \theta_0 + \varphi_1, \sigma_\theta)\phi(f; f_0, \sigma_f) \\ - A \cdot \phi(u; d, \sigma_u)\phi(v; 0, \sigma_v)\psi(\theta; \theta_0 + \varphi_2, \sigma_\theta)\phi(f; f_0, \sigma_f) \end{aligned}$$

where the orientation changes φ_1 and φ_2 are defined as $\varphi_1 = \arctan\left(\frac{v}{\rho - u}\right)$ and $\varphi_2 = \arctan\left(\frac{v}{-\rho + d - u}\right)$. The excitatory pattern is similar to the previous cases except that the peak local orientation at any position (u, v) is always oriented toward the same point $(\rho, 0)$, thus expressing convergence of local orientations. As a special case, this leads to iso-orientation when ρ goes to infinity. We later use the value ρ/σ_v (the converging distance relative to the envelope width) to classify the excitation pattern into iso-oriented excitation ($\rho/\sigma_v > 10$; Fig. 1*C*,*d*) or orientation-convergent excitation ($\rho/\sigma_v \leq 10$; Fig. 1*C*,*e*). The inhibitory pattern is mirror symmetric to the excitatory pattern with the center shifted by distance d along the envelope.

Using these descriptive functions, each unit was classified into the following steps. (1) The basis vector of the unit was fitted with the above four functions. To constrain an estimated envelope from overly exceeding the *x*–*y* boundaries, additional data points of zero values were padded right outside the boundaries. (The zero values were used only for the constraint and thus removed after fitting.) If fitting with the functions of Type II or IV resulted in the center of the inhibitory pattern exceeding the boundary, then these functions were not considered. Also, we considered only the functions that gave a goodness-of-fit satisfying $R^2 \geq 0.5$ (which ensures statistical significance with $p < 10^{-100}$ in F test comparing to fitting with the constant zero function); if all functions gave a bad fit, then the unit was discarded. (2) The inhibition type was determined by the best-fitting function (since the inhibition type was unambiguously defined for each of the four functions). To accomplish this with rigorous statistical criteria, we used the Akaike information criteria, $A_i = 2P_i + N(1 - R_i^2)$, where P_i is the number of free parameters in the ith function ($P_1 = P_2 = P_3 = 10$ and $P_4 = 11$), N is the number of data points (i.e., Layer 2 units; $N = 1296$), and R_i^2 is the goodness-of-fit value for the ith function. Then, we selected the function j that gave a significantly better fit than any other function in the sense that the relative likelihood $\exp\left(\frac{A_j - A_i}{2}\right) < 0.05/3$ was for all $i \neq j$ (Bonferroni correction for three comparisons). If no function satisfied this criterion, then the unit was discarded. (3) The excitation type was orientation convergent if the best-fitting function was Type IV and $\rho/\sigma_v < 10$. Otherwise, the excitation type was iso oriented.

We call a unit well classified if it was not discarded in step 1 or step 2. Further analysis was conducted only on such well classified units.

Preferred natural image patches. To investigate what kind of features in input images could be detected by each unit, we presented a set of 100,000 natural image patches and visualized the most preferred patches. These input patches were extracted from the image dataset separately from the training set. The extraction method was similar to the training data except that only high-contrast patches were used (variance, >0.89; acceptance rate, 43%) to facilitate visual identification of selected features.

Local orientation organization. The way our model units processed local orientations was compared with macaque V2 cells, following the experimental protocol introduced by Anzai et al. (2007). We measured, for each model unit, the responses to square grating stimuli of size 12 × 12 pixels that were presented at 6 × 6 center locations, 12 orientations, three frequencies, and four phases, in accordance with the Gabor filters in Layer 1. The space orientation response profile was obtained by taking the maximal value at each location and orientation across frequencies and phases; the profile was normalized by the maximum.

To quantify the orientation organization, the set of peak local orientations for each unit was collected from the space orientation profile as follows. At each location, if the local orientation tuning curve was fitted well with a 180° cycled von Mises function ($R^2 > 0.5$), then the peak of

this function was taken; otherwise, if the tuning curve was fitted well with a sum of two von Mises functions ($R^2 > 0.5$), then both peaks of this function were taken; if neither fitted well, then this location was unused. Finally, all of the peaks at a magnitude >0.5 were pooled. After this, the maximal difference and the set of all pairwise differences between the pooled peak orientations were calculated. Comparison of the distributions of these values across the population of units to the experimental result (Anzai et al., 2007) is presented in Results. Note that Anzai et al. (2007), in addition to the experiment using single grating patches, conducted another experiment using double grating patches; a simulation of the latter experiment is also discussed in Results.

Angle selectivities. The way our model units processed angle features was also compared with macaque V2 cells, following the method used by Ito and Komatsu (2004). We used the set of 66 angle stimuli each composed of two line segments of 15 pixels emanating from a given center location toward one of 12 directions at 30° intervals. We measured, for each model unit, the responses to these stimuli at 13 × 13 center locations (at intervals of 2 pixels) and at 0° and 15° rotations. The response profile at the location and rotation where the mean response to all 66 stimuli was maximal was used for further analysis; the profile was normalized by the maximum. The profile was formatted in the upper half of a 12 × 12 matrix, where the column or row corresponded to the direction of one of the two angle components (see Fig. 7*D*); note that, by definition, the profile would look mirror symmetric with respect to the diagonal if the lower half were also shown.

Analysis of the response profile proceeded as follows. First, local maxima were obtained from the response profile after smoothing it with a 3 × 3 Gaussian filter and removing the responses below a threshold of 0.8. The top two local maxima (or the unique maximum if only one existed) were used and the corresponding angles were called peak angles. (If no peak angle was found, then the unit was discarded.) Then, for each peak angle, the number of matrix entries in each "elongation" was counted from the original response profile. There are four types of elongation, as follows: (1) horizontal; (2) vertical; (3) angle; and (4) orientation elongations (see Fig. 7*D*). The horizontal elongation includes the entries above a threshold of 0.6 corresponding to the stimuli sharing one component of the peak angle. The vertical elongation was similar except that it was obtained for the stimuli sharing the other component of the peak angle. The angle elongation included the entries above 0.6 corresponding to the stimuli sharing the same angle width as the peak angle. The orientation elongations included the entries above 0.6 corresponding to the angle stimuli whose bisecting (half-splitting) orientations were equal to or 15° clockwise shifted from that of the peak angle. Comparison of the distribution of peak angle widths and the distributions of the four types of elongations to the experimental result (Ito and Komatsu, 2004) is presented in Results.

Length and width tuning. The length and width tuning patterns of our model units were compared with the data reported for macaque V2 cells (Schmid et al., 2014). That study adopted the experimental protocol described by DeAngelis et al. (1994); we thus simulated the same protocol here. Concretely, we used the set of rectangular patches of grating, with sizes varying from 6 × 6 to 24 × 24 pixels (at intervals of 2 pixels). We measured the responses to these stimuli, for each model unit, at 6 × 6 center locations (at intervals of 4 pixels), 12 orientations, three frequencies, and four phases (in accordance with the Layer 1 layout); in the sequel, we used the mean responses over phases. We determined the preferred center location, orientation, and frequency as those giving the maximal average response for the set of rectangular patches of the smallest size (6 × 6 pixels). The width response profile was defined as the mean responses to the gratings with the optimal length and varied widths that were presented at the preferred center location, orientation, and frequency; the length response profile was defined analogously.

To quantify the tuning patterns, each width or length response profile was fitted with a half-rectified difference of two error functions (integrals of Gaussian function). Then, the width or length giving the peak response (R_{peak}) was obtained from the fitting function. The width or length suppression index was defined as $(R_{\mathrm{peak}} - R_{\mathrm{lim}})/R_{\mathrm{peak}}$, where R_{lim} was the response at the largest width or length in the fitting function. Thus, the suppression index gives the ratio of the maximal response decrease to the peak response. In the study by Schmid et al. (2014), a response profile was also fitted with a single error function (representing a nonsuppressive tuning curve) and the better fitting function was chosen according to a certain statistical criterion. However, we did not adopt this method since our response data were deterministic and a difference of two error functions seemed to reasonably represent both suppressive and nonsuppressive tuning curves. (This modification makes a difference only in the case of a very weakly suppressive tuning curve, for which our approach gives a very small index while the original approach would give zero.) Comparison of the joint distribution of width and length suppression indices to the experimental data (Schmid et al., 2014) is presented in Results. Note that, in addition to responses to rectangular patches of grating, Schmid et al. (2014) analyzed cell responses in an experiment using "orientation-discontinuity" stimuli; the latter was not simulated here since the property that they discovered was not expected to be found in our model (see Results).

Results

Hierarchical model of natural images

In this study, we hypothesized that V2 may perform sparse coding of V1 outputs. Accordingly, we constructed a three-layer hierarchical network model, as illustrated in Figure 1*A*. Layer 1 (model V1 simple cell layer) was a pre-fixed bank of Gabor filters of all combinations of 6 × 6 center locations, 12 orientations, three frequencies, and two phases. Layer 2 (model V1 complex cell layer) implemented standard energy models of complex cells by integrating the outputs of each pair of Layer 1 units at the same center location, orientation, and frequency, but with different phases. Layer 3 (model V2 layer) used sparse coding of Layer 2 outputs with 800 units. Note that, although we followed the generally accepted view that V1 has two types of cells, phase-sensitive simple cells and phase-insensitive complex cells, and V2 receives major projections from V1, the model was certainly a radical simplification and was not meant to be an accurate descriptive model. The filter weights between Layers 2 and 3 were subject to learning with natural image patches (after processing by Layers 1 and 2) using a sparse coding principle. Our specific learning method was a combination of strong dimension reduction for spatial pooling and overcomplete ICA for feature extraction. Natural image patches were extracted from the ImageNet10K dataset (Deng et al., 2010). The dataset was huge in terms of size and variety: it contains 9,000,000 images of 10,000 categories, including natural and urban scenes, natural and artificial objects, humans, and animals (for details of the model construction, see Materials and Methods).

To visualize the internal representations of the model V2 units after learning, we use the following two formats: the filter format and the basis format. In the filter format, we show the vector associated with each unit in the filter weight matrix estimated by the learning method. In the basis format, we show the vector associated with the unit in the network interpreted as a generative model (the basis vectors can be obtained essentially by inverting the filter weight matrix). In both formats, each model V2 unit is drawn as a set of ellipses each representing a model V1 complex unit with the indicated center location, orientation, and inverse of frequency (size), where the color shows the normalized (filter or basis) weight value between the model V1 complex unit and the model V2 unit. For concise presentation, we show only the maximum positive and the minimum negative weights as each position. (Despite the drastic omission, this visualization does not lose much information in the internal representation (for visualization details, see Fig. 1*B*; also see Materials and Methods.) Below, we mainly use the basis format for informal illustration and formal classification since this format exposes the intuitive

meaning of each model unit; we later use the filter format when describing details of the response properties.

Figure 2 shows 40 example units in the basis format, illustrating the regularity and variety of the model units. All shown units have localized excitatory subfields and some also have inhibitory subfields. For many units, each excitatory or inhibitory subfield integrates rather similar local orientations. However, the detailed structures of the excitatory and inhibitory subfields seem to reflect distinct categories. The units in Figure 2, column a, have an excitatory subfield with similar local orientations, together with a weak, broad inhibitory pattern; we call this type of unit "iso-oriented excitation with broad inhibition." The units in Figure 2, columns b–d, also have an iso-oriented excitatory subfield, but with a symmetric inhibitory subfield appearing on the side (Fig. 2, column b), in a cross-like formation (Fig. 2, column c), and at the end (Fig. 2, column d); we call these types of unit "iso-oriented excitation with side, cross, or end inhibition," respectively. The units in Figure 2, column e, have a localized excitatory subfield in which the combined local orientations all seemingly converge to a certain point; a symmetric inhibitory subfield appears at the end; we call this type of unit "orientation-convergent excitation with end inhibition." An additional observation, although not a focus here, is the variety in frequency integration. For example, some units like those in Figure 2, column a, rows 2 and 5, have prominent frequencies higher than other units like those in Figure 2, columns a and c, row 1. Also, while many units integrate similar frequencies uniformly, some units combine higher frequencies in some parts and lower ones in other parts (e.g., Figure 2, column b, row 3, and column e, row 1).

We have exemplified in Figure 2 five types of model units, which were in fact representatives of the entire V2 model. (In particular, we rarely found a unit with orientation-convergent excitation in conjunction with broad, side, or cross inhibition.) To quantify this, we introduced a formal classification method for the model units. The classification used four algebraic functions for describing basis representations of the four inhibition types (broad, side, cross, and end inhibition types). Each function was defined as a four-dimensional Gaussian-like function (x and y coordinates, orientation, and frequency) or a difference of two such functions to describe the shapes of the excitatory and inhibitory subfields. For the end inhibition type, however, the local orientations in either excitatory or inhibitory subfield were allowed to converge to a certain point, where the convergence was steeper if the distance from the subfield center to the converging point was shorter. Thus, the inhibition type of each unit was determined by the best-fitting function, and the excitation type was then determined using the parameters controlling the degree of convergence (for details, see Materials and Methods; Fig. 1*C*).

Figure 3 summarizes the result of the formal classification. For 93.7% of the units, one of the four algebraic functions gave a fit that was good by itself and significantly better than any other functions (see Materials and Methods). We call such units well classified; further analysis was conducted only on those units. Figure 3*A* shows the distribution of the goodness-of-fit values (R^2) for the best fitted functions (the filled and the unfilled bars indicate the well classified and the unclassified units, respectively). The example units shown in Figure 2 were actually randomly selected from each class formally defined in this way. (The units are sorted by goodness of fit; those in the bottom are therefore around the borderlines between different classes.) Among the five types, the iso-oriented excitation with broad inhibition type (~30%), the orientation-convergent excitation with end-inhibition type (~23%), and the iso-oriented excitation with side inhibition type (~21%) were relatively more frequent, though the other types were also quite common (Fig. 3*B*). It is quite noteworthy that units with such a complex structure as orientation-convergent excitation emerged so frequently.

To what features in natural images do these excitation and inhibition patterns of model units correspond? To gain an intuitive understanding, we input randomly selected 100,000 natural image patches to the model and visualized the most preferred patches of each model unit. Figure 4 shows the seven most preferred patches of 15 model units given in the top three rows in Figure 2. The units seen in Figure 4, rows a1–a3, b2, b3, c1–c3, and d1–d3, clearly preferred patches containing a contour corresponding to the excitatory pattern. In particular, for the units in Figure 4, rows d1 and d2, with end inhibition, the short contours in the preferred patches often stopped at the point where the inhibition started. The units in Figure 4, rows e1–e3, in a similar way to those in Figure 4, rows d1–d3, preferred patches containing some feature that stopped correspondingly with the end inhibition; the feature was somewhat more complicated than a simple contour and looked more like an acute or round corner. The unit in Figure 4, row a2, in fact preferred patches containing a texture feature reflecting the wider excitatory pattern with multiple parallel orientations. Such texture-selective units were relatively few but not uncommon; Figure 4 shows two additional examples with side inhibition (Fig. 4, row x1) and with end inhibition (Fig. 4, row x2). Although showing all units here is impossible, the unit type generally reflected the preferred feature for units not shown: iso-oriented excitation units preferred contour or texture features and orientation-convergent excitation units preferred corner-like features.

While the basis format provides an intuitive view on what each unit represents, the filter format is important in understanding how each unit responds to external inputs, as discussed in the subsequent section. Figure 5 shows the filter representations of the model units corresponding to the top three rows of Figure 2. For each unit, the basis and filter vectors look reasonably similar in the overall excitatory and inhibitory integration pattern of local orientations, which means that the response properties generally reflect the structure in the basis representation. However, details in these formats are different. The differences are related to the statistical structure in Layer 2 outputs (analogous to the $1/f$ spectrum in natural image statistics), where finer-grained structures (e.g., higher frequencies) have lower magnitudes, and, accordingly, the filters amplify them (see Materials and Methods). This has the following specific effects: (1) the size of the prominent pattern in the filter is often smaller than in the basis; (2) the prominent spatial frequency is sometimes higher in the filter; (3) the filter is somewhat noisier (which would make fitting difficult); (4) some units with broad inhibition in the basis format often have weak side inhibition in the filter format; and (5) in the filter, a strong excitation is often overlaid with a weak orthogonal inhibition, and a strong inhibition is often overlaid with a weak orthogonal excitation. The latter two are particularly important in clarifying some unintuitive details of the response properties shown below.

Homogeneous and heterogeneous integration of local orientations

An experimental study (Anzai et al., 2007) suggested that macaque V2 represents both homogeneous and heterogeneous excitatory integration of local orientations. This property may be explained in our model since it had units of both iso-oriented and orientation-convergent excitation types.

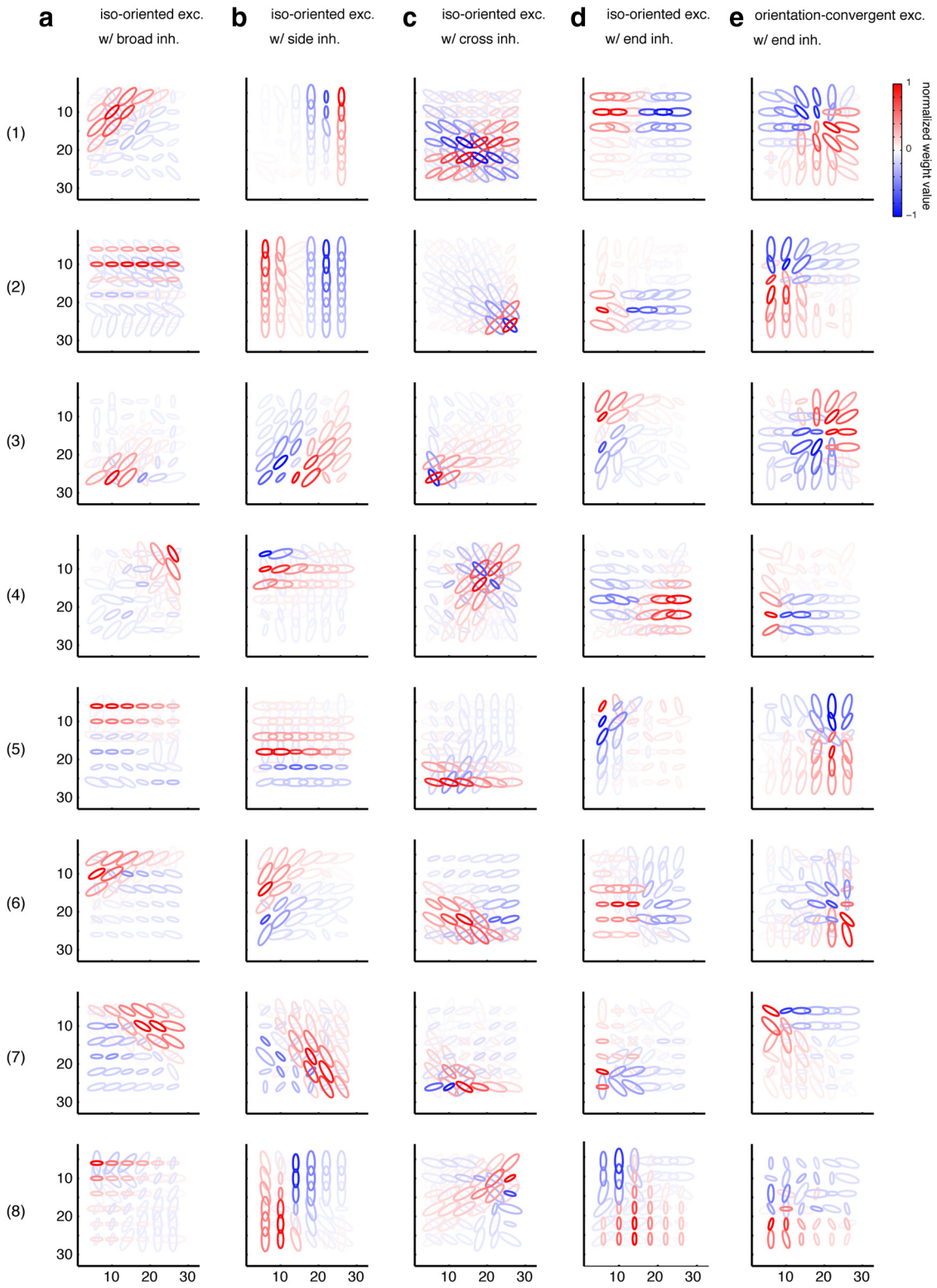

Figure 2. Classified examples of model V2 units in the basis format. Each column shows randomly selected eight units of one of five types, as follows: column ***a***, iso-oriented excitation with broad inhibition; column ***b***, iso-oriented excitation with side inhibition; column ***c***, iso-oriented excitation with cross inhibition; column ***d***, iso-oriented excitation with end inhibition; and column ***e***, orientation-convergent excitation with end inhibition. The units in each column are sorted by goodness of fit (highest in the top row). See Figure 1*B* for the visualization format. exc., Excitation; inh., inhibition.

To reproduce the experimental result, we followed the method used by Anzai et al. (2007), probing the sensitivity of each model V2 unit to local orientations by presenting oriented local grating stimuli at different positions (see Materials and Methods). Figure 6*A* shows the obtained space-orientation response profiles for the units shown in the top three rows in Figure 2, where the local orientation tuning at each position is plotted in polar coordinates (the responses are normalized by the maximal value).

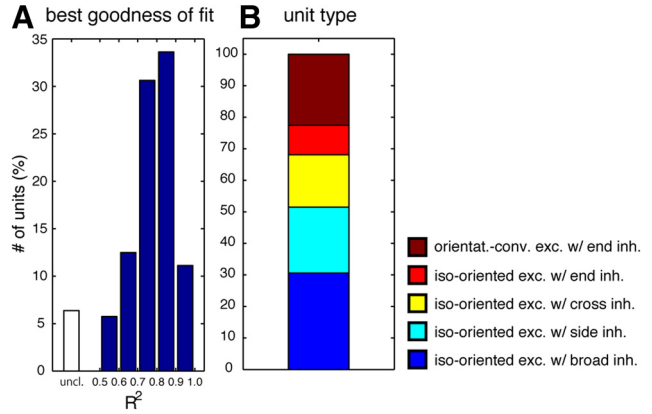

Figure 3. Classification of model V2 units. ***A***, The distribution of the goodness-of-fit values (R^2) for the best-fitting functions. The filled and the unfilled bars indicate the well classified and the unclassified units, respectively. ***B***, The distribution of five types in the model (only for well classified units). The colors of the bars show the unit types (see legend). exc., Excitation; inh., inhibition.

The response profile of each unit generally had parallels with the excitatory part of the basis representation. Most units in Figure 6*A*, columns a–d, which had iso-oriented excitation in the basis representations, had fairly homogeneous orientation integration. On the other hand, the units with orientation-convergent excitation in Figure 6*A*, column e, had somewhat heterogeneous integration, with the orientations gradually changing depending on the position. However, the response profile for the unit in Figure 6*A*, column b, row 1, was rather dissimilar to its basis representation, where orthogonal excitation could not be found. This can be understood from the filter representation shown in Figure 5. In Figure 5, column b, row 1, strong vertical excitation was accompanied by strong vertical inhibition on the side, and this inhibition was overlaid with weak horizontal excitation. The responses to horizontal orientations in the response profile were caused by this weak horizontal excitation, but were somewhat exaggerated since the responses to vertical orientations were suppressed by the strong vertical inhibition close to the vertical excitation. (Many other side inhibition units in fact had similar heterogeneous space-orientation maps, which could be explained in the same manner.) The response profiles shown in Figure 6*A* are qualitatively similar to the examples given by Anzai et al. (2007); our homogeneous units [Fig. 6*A*, columns a (rows 1–3), b (rows 2, 3), c (rows 1–3), d (rows 1–3)] were similar to those in their Figure 1*a*, our heterogeneous unit [Fig. 6*A*, column b (row 1)] was similar to that in their Figure 1*b*, and our heterogeneous unit [Fig. 6*A*, column e (rows 1–3)] was similar to that in their Figure 1*d*.

As in the experimental study, we quantified the degree of orientation heterogeneity of each unit by calculating the maximal difference between the peak orientations at different positions (see Materials and Methods). Figure 6*B* shows the distribution of the maximal orientation differences of the entire population in the model (bars). Clearly, the distribution has two prominent peaks around 0° and 90°, which is consistent with the experimental result (Anzai et al., 2007) with similar peaks (Fig. 6*B*, magenta curve). Figure 6*B* also shows the proportions of the unit classes within each bin, giving one interpretation to the experimental data. Note that approximately one-half of the most heterogeneous units (75–90°) could be explained by the orientation-convergent excitation type of units, so that the peak at ~90° would be much less pronounced if these units were not present. Since such a peak did not exist for V1 according to the same experimental study, units of orientation-convergent excitation type might play a key role in producing the V2-specific response properties.

Note that, if a unit had a completely random orientation organization, then the maximal orientation difference would be close to 90°. Figure 6*C* excludes this possibility. That is, if the orientation organization were completely random, the distribution of all pairwise orientation differences (the differences between all pairs of peak orientations) pooled across the population of heterogeneous units (maximal orientation differences >30°) would be uniform. However, the actual distribution (bars) of all pairwise orientation differences shows prominent peaks at ~0° and 90°. This result is also consistent with the experimental data (Anzai et al., 2007; Fig. 6*C*, magenta curve).

In addition to the above experiment, Anzai et al. (2007) conducted another experiment using a pair of grating stimuli presented at different locations. This experiment could reveal orientation interaction at two locations as a modification of the orientation tuning curve at one location, induced by the orientation at the other location. We also simulated this experiment in our model and quantitatively compared the orientation interactions. We considered four types of orientation interactions based on the tuning similarity index (TSI) that was proposed by Anzai et al. (2007) for measuring the strength of nonlinearity in the orientation interaction. The four types were as follows: (1) additive (vertical shift of tuning curve; TSI > 0.8); (2) multiplicative (amplitude reduction; 0.8 ≥ TSI > 0.2); (3) flattening (loss of tuning; 0.2 ≥ TSI > 0); and (4) inverting (upside-down tuning curve; TSI ≤ 0). In our model V2, we found units that had interactions of types 1–3, where type 1 was far more frequent than the other types; we found no unit with interactions of type 4 (data not shown). This result is not surprising since our model V2 units are simple linear–nonlinear models taking V1 outputs. In contrast, in macaque V2, Anzai et al. (2007) found cells with interactions of all types. In particular, interactions of type 2 were more frequent than the other types. Also, a relatively small population of interactions of type 4 existed. This suggests that a part of the actual V2 neurons may involve a strong nonlinear computation beyond our simple sparse coding model, and an additional mechanism must be introduced to account for such behavior (see Discussion).

Angle selectivities

Macaque V2 is also known to exhibit various angle selectivities (Ito and Komatsu, 2004). This property also may be explained in our model since it had units of various orientation heterogeneities.

We simulated, in our model, the protocol used in the experimental study (Ito and Komatsu, 2004) by measuring the responses of each model V2 unit to a set of 66 angle stimuli, each of which was composed of two end-adjoined line segments emanating in two directions (from 12 directions at 30° intervals; see Materials and Methods). Figure 7*A* shows the response profiles of the units given in the top three rows in Figure 2. Each response profile arranges the angle stimuli in the upper half of a matrix in such a way that the row or column corresponds to one of the 12 directions composing the angle stimuli (Fig. 7*D*); the darkness of each displayed stimulus indicates the normalized responses. Note that the top edge and the right edge of the matrix are conceptually continuous; the fourth element on the top row is next to the fourth element on the right-most column, for example. The red boxes indicate the peak angle stimuli, where at most two of them were determined for each unit based on its responses.

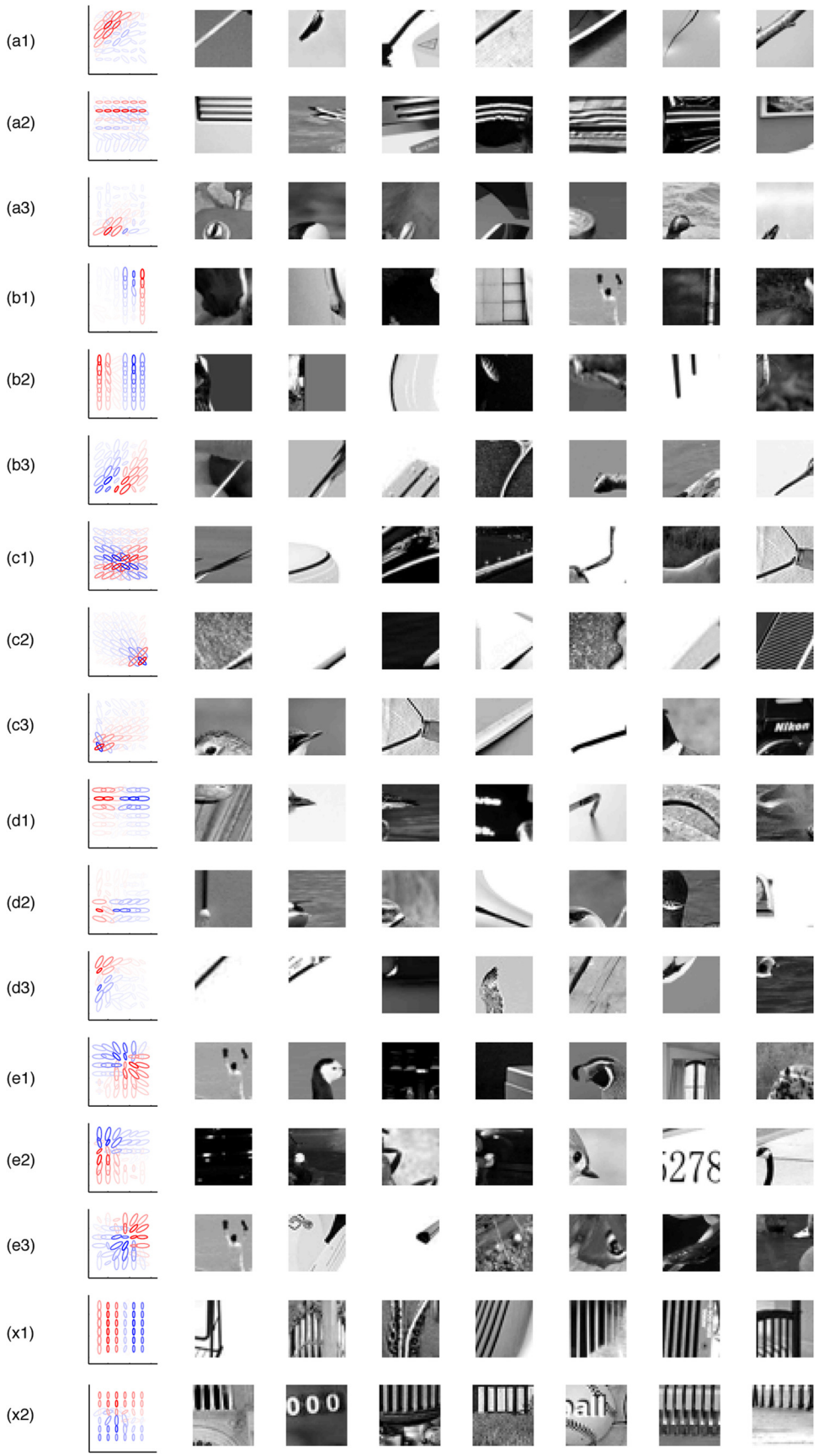

Figure 4. Preferred natural image patches of model V2 units. Each row shows the basis vector of a unit (in the left-most panel in the same format as Figure 1*B*) and its seven most preferred natural image patches. The units shown here correspond to the top three rows in Figure 2 (using the column and row labels there), in addition to two extra example units (x1 and x2) with wide excitatory subfields.

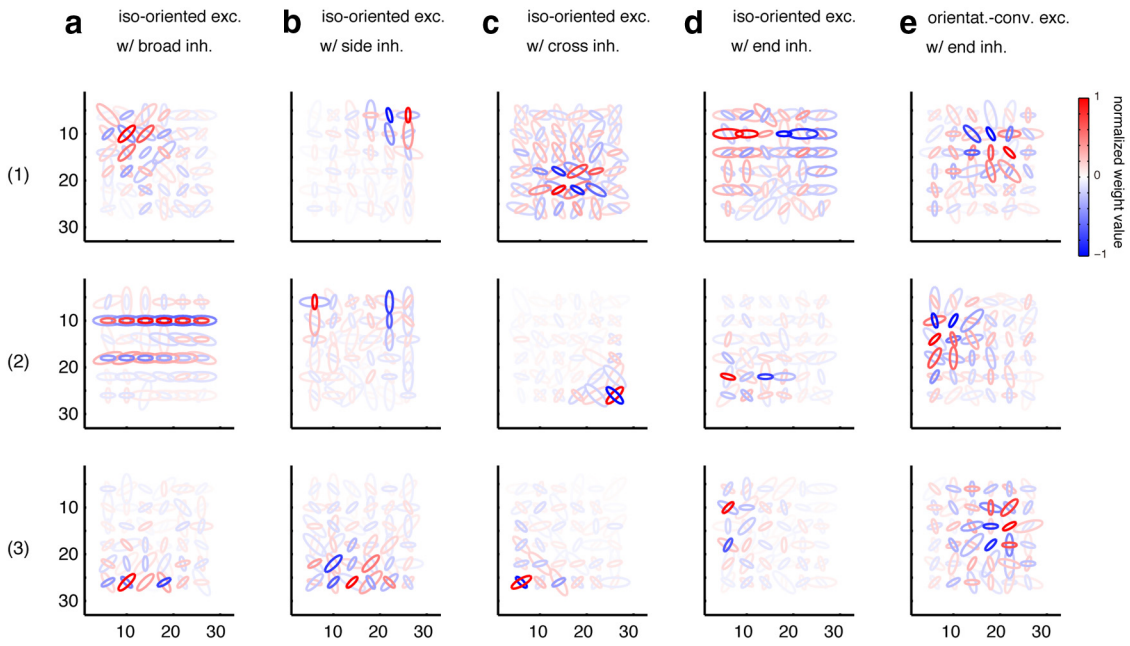

Figure 5. Example model V2 units in the filter format. The units shown correspond to those in the top three rows in Figure 2. (The column and row labels correspond to those in Figure 2.) See Figure 1*B* for the visualization format. exc., Excitation; inh., inhibition.

Two observations can be made from these examples. First, the peak angles for each unit looked approximately similar to the corresponding basis representation. Since the units in Figure 7*A*, columns a–c, had iso-oriented excitatory subfields, many of these preferred wide angles such as 150° and 180°. However, there were a few units in these columns that preferred very narrow angles such as 30° (e.g., units in Figure 7*A*, columns a and c, row 3). This happened since those units integrated not only exactly equal orientations but also neighboring orientations. The units in Figure 7*A*, columns d, e, preferred 30° angles, which is reasonable not only because the basis representations looked quite like sharp angles, but also because the end inhibition in those units prevented them from responding strongly to more elongated stimuli. Second, the response level tended to be somewhat retained for stimuli close to the peak stimulus. In particular, many units had an elongated area around the peak angle, retaining almost the peak level either in the same row or in the same column [e.g., units in Figure 7*A*, columns a (row 2), c (row 1)], which means that, of the two components of the peak angle, the unit was more sensitive to one than the other.

Figure 7*B* shows the distribution of angle widths for the entire population in the model (bars). A wide range of selectivities to angle widths was represented, with prominent peaks at 30° and 180°, which were consistent with the experimental data (Ito and Komatsu, 2004; Fig. 7*B*, magenta curve). Figure 7*B* also shows the proportions of the unit classes within each bin. In particular, the units preferring 30° angles included both iso-oriented and orientation-convergent excitation types, which could mean that selectivities to sharp angles and flat angles might be difficult to distinguish by using such simplistic stimuli. A follow-up article (Ito and Goda, 2011) of the experimental study also indicated a possible discrepancy between preferences of 30° angles and the actual representations.

Figure 7*C* shows the distributions of elongations of angle tuning in the model (bars). Here, elongation is defined as the number of elements above a certain threshold of activation in the matrix that share a certain property with the peak angle. There are four types (Fig. 7*D*). Primary and secondary elongations correspond to the horizontal and vertical axes in the matrix and count the number of elements sharing one of the two components of the peak angle; primary elongation refers to the larger one. Angle elongation corresponds to the left-up-to-right-down axis and counts the number of elements with the same angle width as the peak angle. Orientation elongation corresponds to the left-down-to-right-up axis and counts the number of elements where the bisecting orientation of the angle is similar to the peak angle (see Materials and Methods). As evident from Figure 7*C*, the distribution of primary elongations was much broader than the distributions of the other types of elongations, indicating the biased sensitivity to one component of the peak angle across population, which, again, is consistent with the experimental data (Ito and Komatsu, 2004; Fig. 7*C*, magenta curves).

Length and width suppression

Yet another property known in macaque V2 is suppression of responses to grating stimuli when they are lengthened or widened to some extent (Schmid et al., 2014). This property may also be explained since our model units have a variety of inhibition patterns.

We simulated the experimental method used by DeAngelis et al. (1994), which was adopted by Schmid et al. (2014). We thus presented, to the model, optimally oriented grating stimuli with rectangular envelopes of varied lengths and width (the orientation of the envelope and that of the grating were aligned; see Materials and Methods). Figure 8*A* shows pairs of the tuning functions for varied lengths with the optimal width (left) and for varied widths with the optimal length (right) for each unit given in the top three rows in Figure 2.

Overall, the tuning functions reflected the inhibition patterns in the basis representations of these units. For the side inhibition units in Figure 8*A*, column b, the response was monotonically increased when the grating stimulus was lengthened, while suppressed when the stimulus was widened. For the end inhibition units in Figure 8*A*, columns d, e, the response was increased when

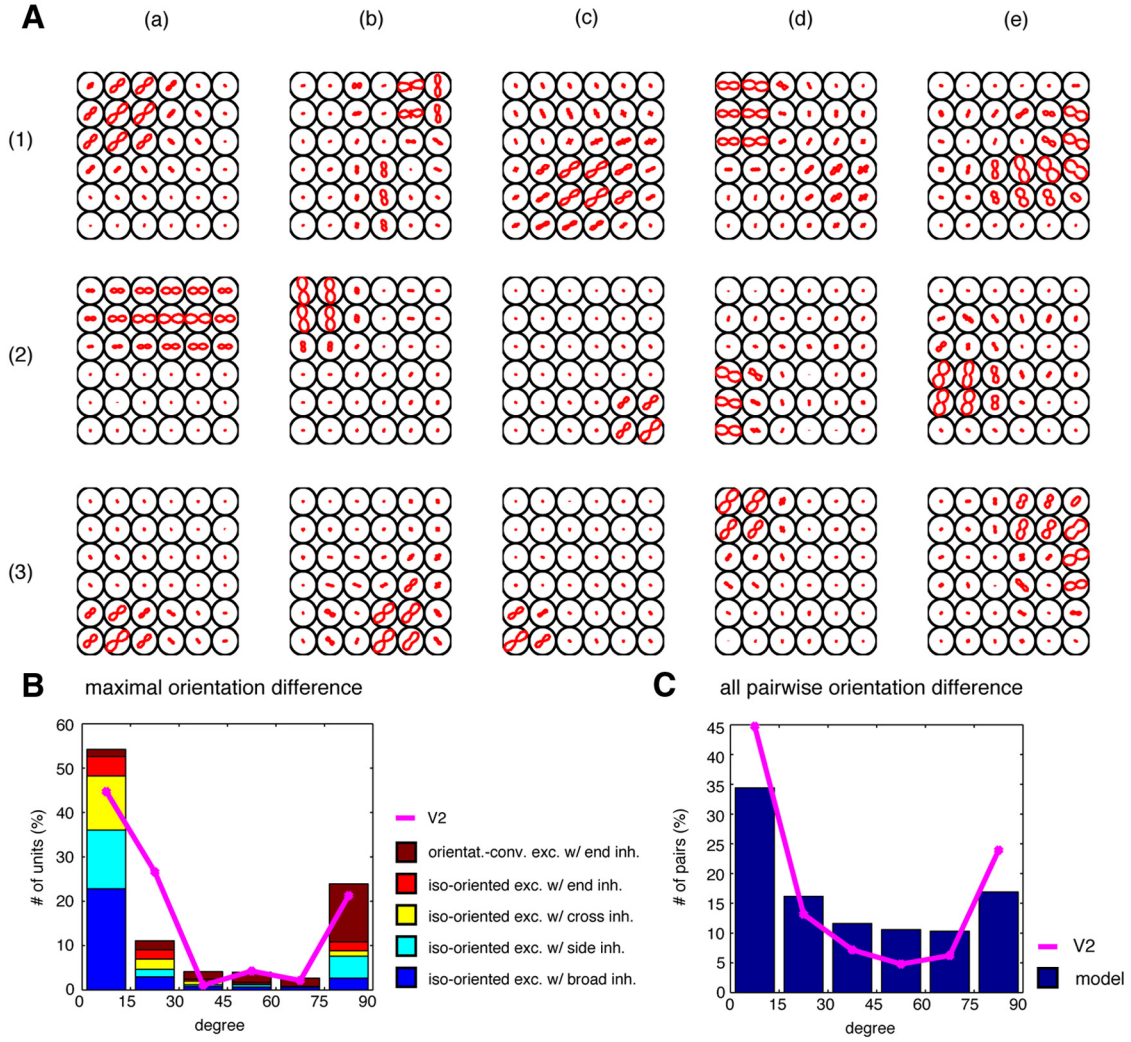

Figure 6. Local orientation integration property of model V2 units compared with the macaque V2 data (Anzai et al., 2007). ***A***, The space-orientation response profiles of the units given in the top three rows in Figure 2 (using the same labeling convention). For each unit, the local orientation tuning at each position is plotted in polar coordinates. ***B***, The distribution of maximal orientation differences of all model units, with the proportion of each unit class in each bin (bars). The plot is overlaid with the replotted analogous distribution reported for V2 (Anzai et al., 2007) (magenta curve). ***C***, The distribution of all pairwise orientation differences pooled across all heterogeneous units (whose maximal orientation differences are $>30°$; bars), overlaid with the replotted analogous distribution (Anzai et al., 2007) in magenta curve. The replots in ***B*** and ***C*** combine the bins for both positive and negative orientation differences with the same magnitude in the experimental results shown in Anzai et al. (2007; their Fig. 2*b*,*f*). exc., Excitation; inh., inhibition.

the stimulus was widened, while suppressed when the stimulus was lengthened. The broad inhibition units in Figure 8*A*, column a, and the cross inhibition units in Figure 8*A*, column c, generally had little suppression in either tuning function, except that some units in Figure 8*A*, column a, had moderate width suppression (e.g., Fig. 8*A*, column a, rows 1, 2), because they in fact had weak side inhibition in their filters (Fig. 5).

The degree of suppression in each length or width tuning function was quantified by a suppression index ranging between 0.0 and 1.0, where a larger value indicated a stronger suppression. [Although a slightly different definition of suppression index from the experimental study by Schmid et al. (2014) was used due to the nature of data, it made a difference only for very small suppression indices (see Materials and Methods).] Figure 8*B* plots the joint distribution of the length and width suppression indices for all units, where the color shows the unit type. The distribution was clearly bimodal, where the majority of units had large values for either the length or the width suppression index, not both, reflecting their inhibition types. That is, most units with large length suppression indices were of the end inhibition type; those with large width suppression indices were of the side inhibition type; those with relatively small values for both suppression indices were of either the broad inhibition or cross-inhibition type. The medium width suppression indices for most broad inhibition units were due to weak side inhibition in their filters. Compare this simulated result with the corresponding experimental data (Schmid et al., 2014) in Figure 8*C*. The distribution from the experiment also had two modes with large values for, exclusively, either the length or the width suppression index, which is consistent with the simulation result. The same experimental study showed that the joint distribution for V1 had only one mode with correlated length and width suppression indices. Thus, having separate side inhibition and end inhibition units might be crucial for producing the V2-specific response properties. (The values of the suppression indices were generally much larger in the simulated result, which means that the relative strengths of suppressive fields in the model units were not so accurate with respect to the actual neural representations.)

In addition to the dissociated length and width suppression property, Schmid et al. (2014) reported another property that was

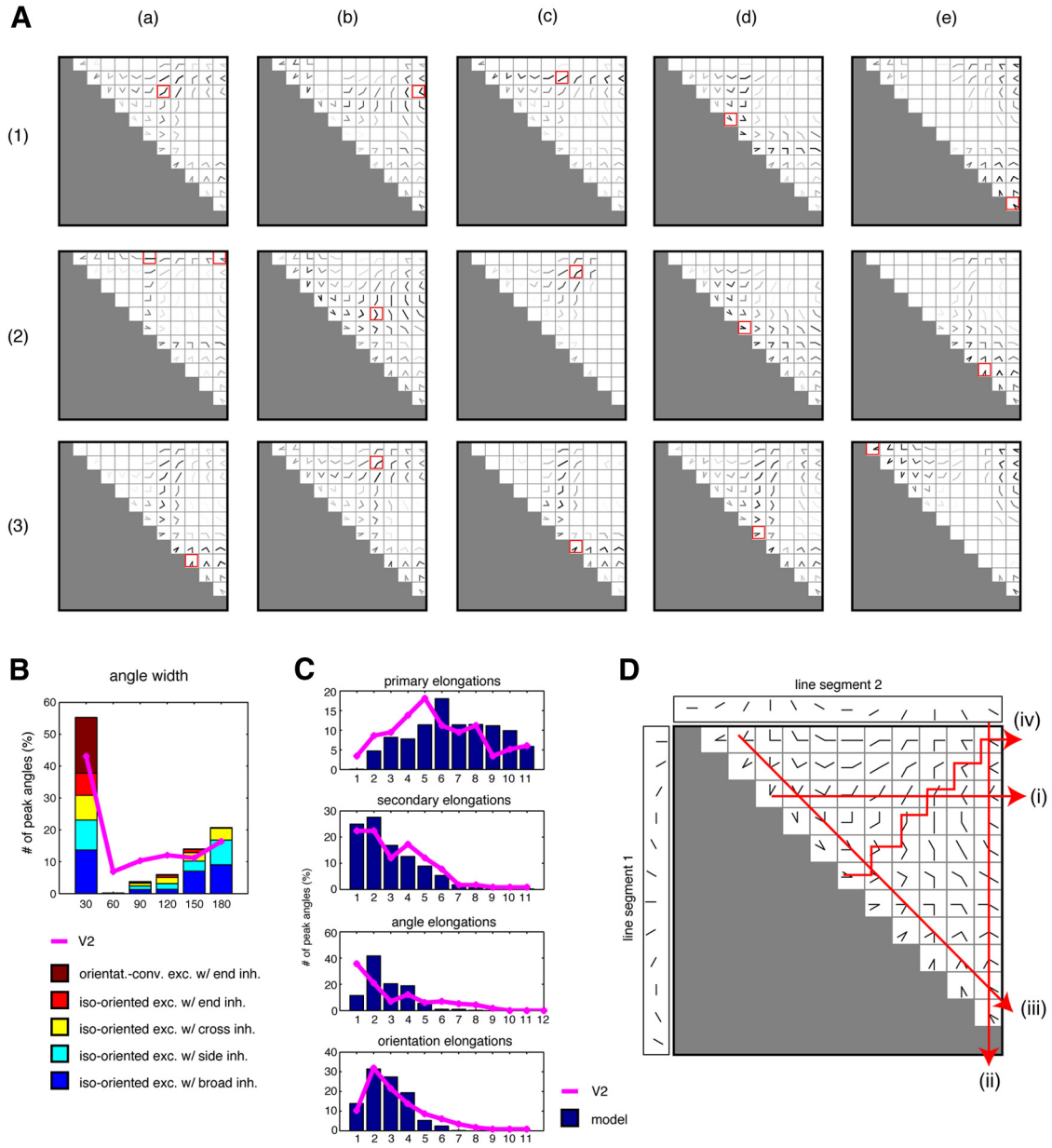

Figure 7. Angle selectivity property of model V2 units compared with the macaque V2 data (Ito and Komatsu, 2004). ***A***, The response profiles for the units in the top three rows in Figure 2 (with the same labeling convention). Darkness of each stimulus indicates the response. The red boxes show the peak angle stimuli. ***B***, The distribution of preferred angle widths with the proportions of unit classes in each bin (bars), overlaid with a replot (magenta curve) of the analogous distribution from the study by Ito and Komatsu (2004). ***C***, The distributions of elongations of peak response areas in primary, secondary, angle, and orientation axes (bars), overlaid with replots (magenta curves) from the study by Ito and Komatsu (2004). ***D***, The matrix format for the 66 angle stimuli composed of two line segments, overlaid with four elongation types: horizontal (i), vertical (ii), angle (iii), and orientation (iv). exc., Excitation; inh., inhibition.

revealed in an experiment using "orientation discontinuity" stimuli (a grid of rectangular regions that are filled with randomly oriented grating patches). However, we did not simulate this experiment since their data analysis crucially depended on the timings and the temporal traces of neural responses (while our model yields only a scalar-valued response), and it did not seem possible that such complicated nonlinear orientation interactions between two positions as were discovered in the experiment would emerge in our simple linear–nonlinear model taking V1 outputs, as discussed in the previous section concerning the simulation of Anzai et al. (2007).

Model variations

To investigate how stable the results shown so far were across various choices of model parameters, we constructed several variations of the model and conducted the same series of analyses. We changed the following three model parameters: (1) the type of image dataset; (2) the number of Layer 3 units; and (3) the number of reduced Layer 2 dimensions. The last parameter was relevant to the strong dimension reduction of the outputs of Layer 2 units performed before computing the overcomplete ICA (see Materials and Methods). In the basic model, these parameters were set to (1) the whole image dataset in ImageNet10K, (2) 800 units, and (3) 100 dimensions. In the variations, we made the following three series of modifications to the basic model: (1) the type of image dataset was changed to a subset of ImageNet10K (images in human, computer, building, or mountain category); (2) the number of Layer 3 units was changed to 200, 400, 1200, or 1600; and (3) the number of reduced Layer 2 dimensions was changed to 200, 400, 600, or 800.

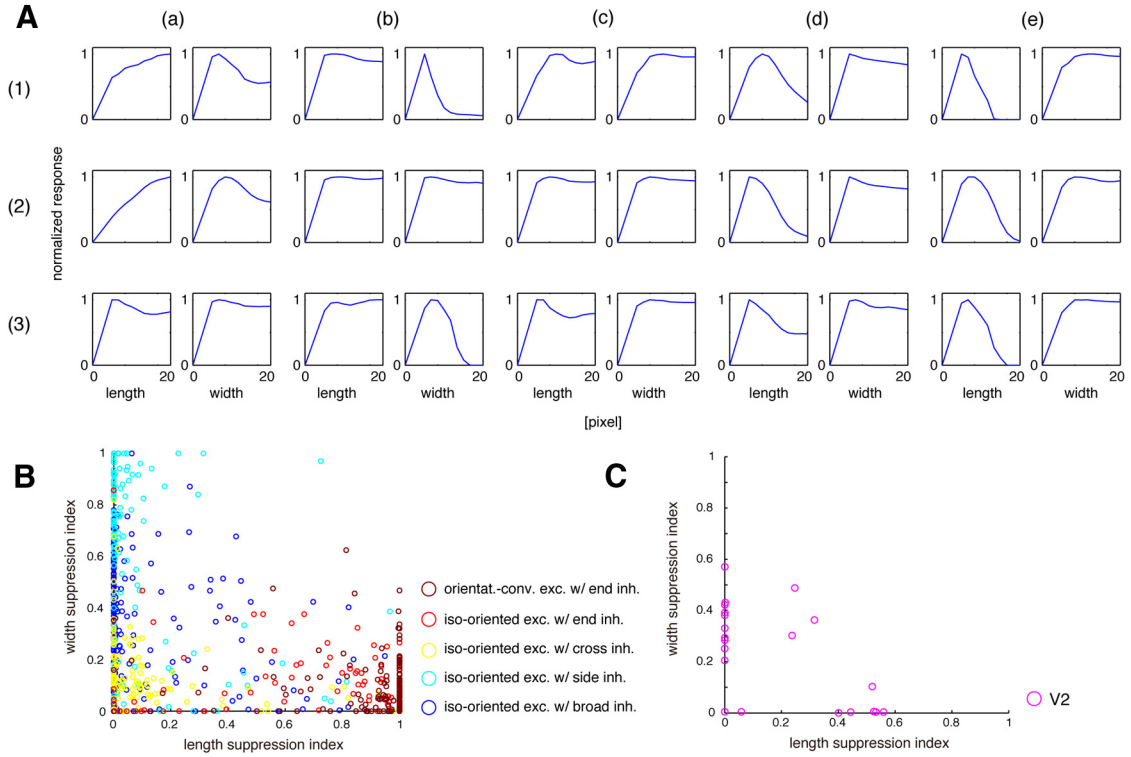

Figure 8. Length and width suppression property of model V2 units compared with the macaque V2 data (Schmid et al., 2014). ***A***, The pair of tuning functions with respect to the width or length (horizontal axis) of a grating stimulus for each unit given in the top three rows in Figure 2 (using the same labeling convention). Responses (vertical axis) are normalized by the maximal value in each tuning function. ***B***, The joint distribution of length and width suppression indices for all units, with colors indicating the unit types. ***C***, Replot of the analogous joint distribution reported by Schmid et al. (2014). A slightly different definition of suppression index was used (see Materials and Methods). exc., Excitation; inh., inhibition.

Figure 9 summarizes the results from these model variations, showing the number of well classified units (Fig. 9*A*), the proportion of each unit type (Fig. 9*B*), the distribution of maximal orientation differences (Fig. 9*C*), the distribution of preferred angle widths (Fig. 9*D*), and the joint distribution of length and width suppression indices (Fig. 9*E*). The population results were overall similar across the model variations, indicating that the learned representations generally had little dependence on a particular choice of the varied model parameters.

The only notable instability was the drop in the number of well classified units and the increase in the ratio of broad inhibition units when the number of reduced dimensions was increased. This was in fact due to the effect of spatial pooling by the strong dimension reduction (see Materials and Methods). To illustrate this, Figure 10 shows the basis representations of a randomly selected 20 units, including ill fitted ones from the models with 100, 200, 400, and 800 reduced dimensions. One can clearly perceive the gradual decrease of the overall sizes of the learned features. In particular, in the case of reduced dimensions 400 (Fig. 10*C*) or 800 (Fig. 10*D*), a large number of units had almost the same sizes as a single complex cell. To quantify this observation, we measured the ratio of the receptive field size of each unit in Layer 3 to the average receptive field size in Layer 2 (see Materials and Methods). The mean ratio indeed significantly decreased as the reduced dimensions increased up to 400, but remained similar afterward (Fig. 11). Moreover, the mean ratio was 1.88 in the case of 100 reduced dimensions, which is comparable to the physiological data (compare Freeman and Simoncelli, 2011, their Fig. 1), thus justifying our choice of the number of reduced dimensions. Figure 10 also shows that the models with 400 and 800 reduced dimensions contained some units whose excitatory and inhibitory subfields were overlapped in a complicated way. These units could in fact be interpreted as side inhibition or end inhibition units where the distance between the excitation and the inhibition was extremely condensed. Our algebraic functions for unit classification were not designed to accommodate such pathological cases and often misclassified those to the broad inhibition type. However, we did not pursue this issue further since those models with higher reduced dimensions were already inappropriate as a V2 model since their receptive field size ratios were too small compared with actual V2.

Discussion

In this article, we investigated a hierarchical statistical model that performed sparse coding of outputs from a standard V1 model. After training with a wide variety of natural image patches, the model units represented five types of excitatory and inhibitory patterns, namely, iso-oriented excitation combined with broad, side, cross, or end inhibition, as well as orientation-convergent excitation combined with end inhibition. Furthermore, the model simultaneously reproduced three kinds of response properties of macaque V2 neurons that were reported in separate experimental studies, namely, local orientation integration (Anzai et al., 2007), angle selectivities (Ito and Komatsu, 2004), and length and width suppression (Schmid et al., 2014). The reproducibility was qualitative, quantitative, and stable across model variations. These results are consistent with the idea that neural representations in V2 use a sparse code of natural images.

We formally classified the model units by fitting their internal representations with our algebraic functions. Relating the classification result with the response properties, we clarified what aspects of model properties were relevant to the reproduced simulation results,

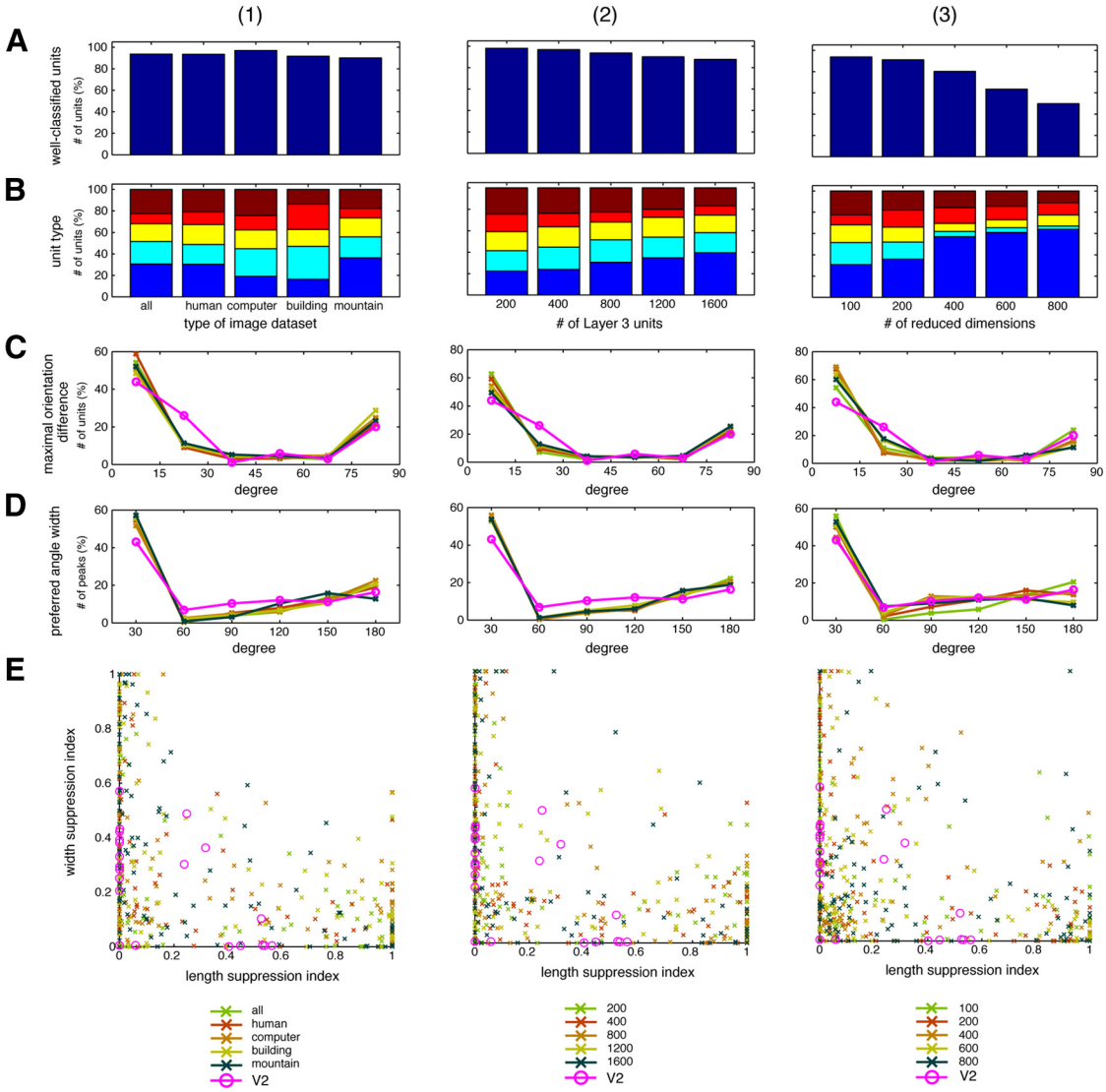

Figure 9. The results for model variations. The varied model parameters were the type of image dataset (1), the number of Layer 3 units (2), and the number of reduced dimensions (3). ***A–E***, The compared properties were the number of well classified units (***A***), the proportion of each unit type (format similar to Fig. 3*B*; ***B***), the distribution of maximal orientation differences (format similar to Fig. 6*B*; ***C***), the distribution of preferred angle widths (format similar to Fig. 7*B*; ***D***), and the joint distribution of length and width suppression indices (format similar to that in Fig. 8*B* except that only 100 randomly selected units are shown for each model variation for readability; ***E***). The colors of curves and points in the plots in ***C***, ***D***, and ***E*** show the model parameter or the corresponding experimental data (each bottom legend).

giving an additional insight into the past experimental data. In particular, the most important type of model unit was orientation-convergent excitation with end inhibition. This type of unit, despite its complex form, was surprisingly frequently found in the model and constituted a majority of all end inhibition units. The reproduction of the aforementioned V2-like tuning properties crucially relied on this type of unit. In particular, these units largely contributed to the population properties with prominent heterogeneities in orientation organization and dissociation of length suppression from width suppression. Since these properties were observed experimentally in V2 but not in V1 (Anzai et al., 2007; Schmid et al., 2014), our model seems to be related to the representations specifically found in V2. (However, it should be noted that precise differences in suppression properties between V1 and V2 are still a matter of debate; Hallum and Movshon, 2014.)

The use of strong dimension reduction before performing overcomplete ICA is a novel aspect in our learning method. For example, in our basic setting, we reduced the dimensions in the outputs of Layer 2 from 1296 to 100 and thereafter estimated 800 independent components. Although this might appear radical and even contradictory at first glance, reducing the dimensions corresponding to low eigenvalues eliminates fine-grained structures in the inputs and thus has an effect of spatial pooling. Indeed, we have observed that the more the dimensions were reduced, the larger the learned features became, while the overall feature shapes remained similar. In particular, when the retained dimensionality was 400 or higher, so many of the learned features became almost identical to a single complex unit that the representations might arguably not be suitable to model V2. This result is related to the fact that the ICA of the result of another ICA essentially only gives an identity matrix; we have also observed that this is often approximately the case even if a nonlinearity like half-rectification is taken in between. Thus, strong dimension reduction might be a simple but powerful method to learn completely new aspects of the data. However, we feel that a deeper and more mathematical understanding of this ap-

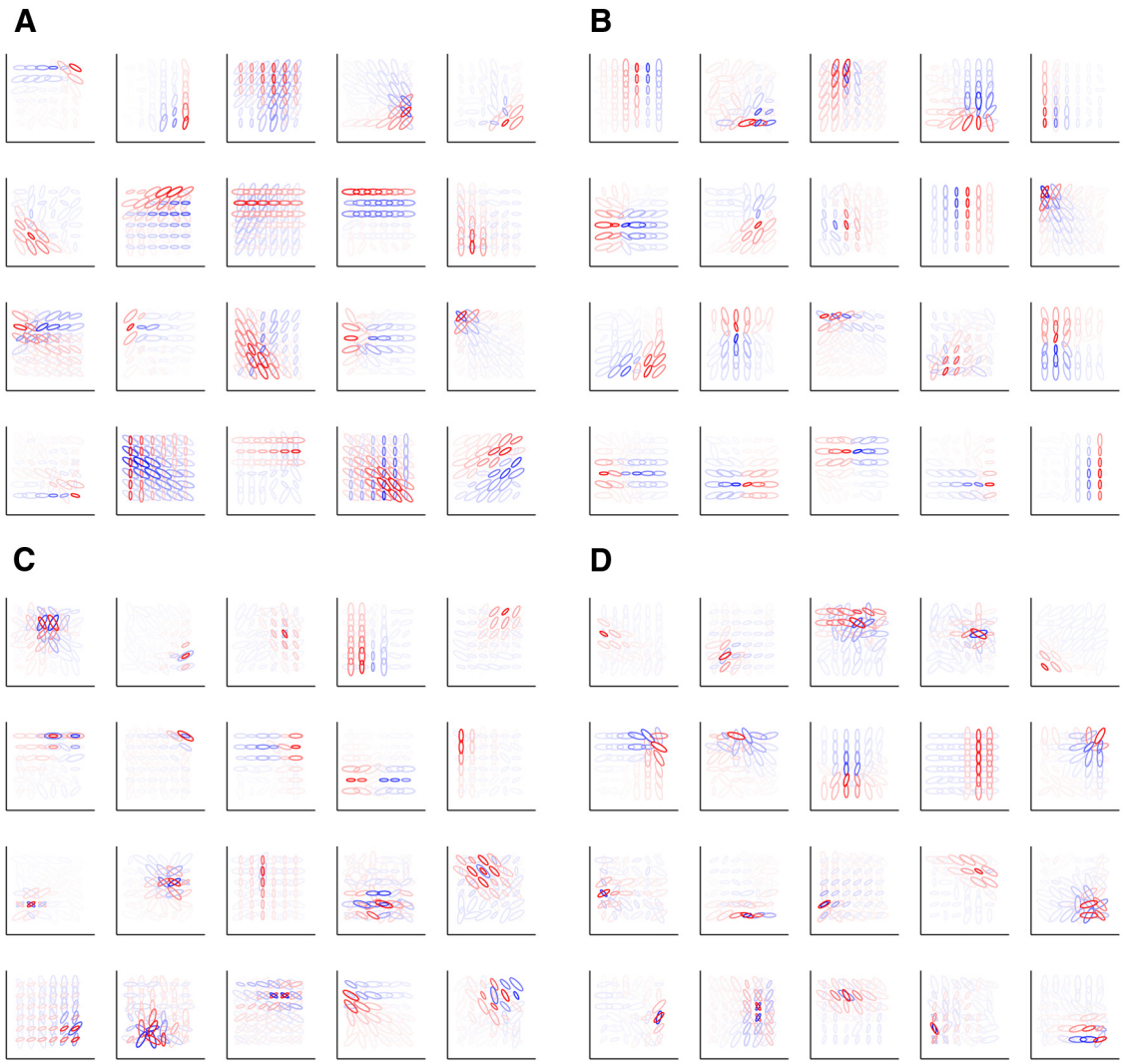

Figure 10. ***A–D***, The basis representations of randomly selected 20 units from each model with 100 (***A***), 200 (***B***), 400 (***C***), or 800 (***D***) reduced Layer 2 dimensions. See Figure 1*B* for the visualization format.

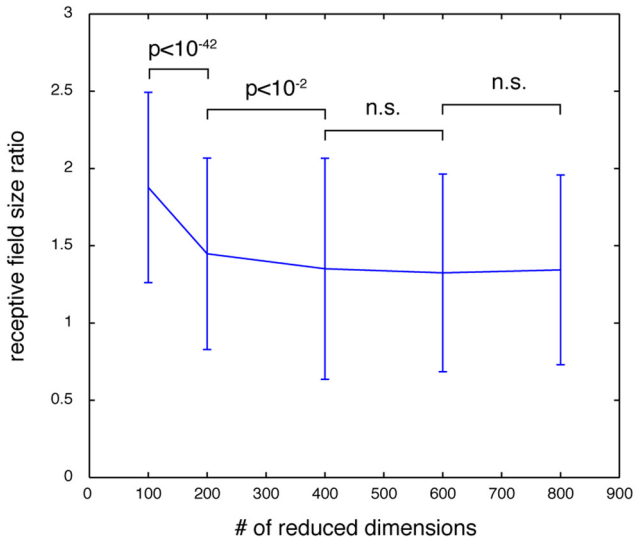

Figure 11. The ratios of receptive field sizes of Layer 3 units to the average receptive field size of Layer 2 units for the models with varied reduced dimensionality. (Each error bar indicates the standard deviation.)

proach is needed beyond the scope of this article, and so it is left for future work.

Our learned model representations were, to some extent, qualitatively similar to filter models of V2 cells estimated in data-driven approaches (Willmore et al., 2010; Tao et al., 2012). In the study by Willmore et al. (2010), filter models assuming a V1-like preprocessing stage with 45° intervals of local orientation detectors were estimated from responses to natural images. They showed that V2 cells often had strong suppression tuned to a particular orientation either equal to or different from the excitatory orientation; they did not particularly report orientation-convergent excitation, but such features would not be detectable with 45° intervals of local orientations. In the study by Tao et al. (2012), filter models were estimated by reverse correlation on local Fourier transforms of dense noise stimuli. They showed that some V2 cells had convergent integrations of local excitatory orientations and that many V2 cells had weak suppressions that were typically orthogonal to the excitation. However, a comparison of the latter study with ours should be made with caution since their method may not estimate filters in the same way as discussed here. [The adequacy of reverse correlation estimation is theoretically not guaranteed for such asymmetric dis-

tribution of inputs (Paninski, 2003); further potential artifacts are discussed elsewhere (Nishimoto et al., 2006).]

Although numerous studies have investigated neural representation in V1 in relation to natural image statistics, only a few focused on V2. Early studies built a sparse coding model assuming a fixed V1-like preprocessing layer, similar to our study (Hoyer and Hyvärinen, 2002; Hyvärinen et al., 2005). However, the resulting representations were much simpler (only with iso-oriented excitation and possible end inhibition) probably due to a lack of spatial pooling mechanisms based on strong dimension reduction. Later studies pursued hierarchical models with all layers learned from natural images, where higher layers estimated variance components (Karklin and Lewicki, 2003; Köster and Hyvärinen, 2010), covariance components (Karklin and Lewicki, 2009), complex-valued sparse coding (Cadieu and Olshausen, 2012), discrete-valued sparse coding (Lee et al., 2008; Hosoya, 2012), or sparse feedforward models (via "noise contrastive" estimation technique; Gutmann and Hyvärinen, 2013). These demonstrated higher representations with more complicated excitatory patterns (elongated and wide subfields) and inhibitory patterns (side, cross, and end inhibitions). None explicitly reported orientation-convergent excitation types combined with end inhibition; in fact, such a structure would be hard to discern from their visualizations since the lower-layer representations were also learned and therefore had an irregular layout. A few of the above studies attempted quantitative comparison with neural properties specific to V2 (Lee et al., 2008; Hosoya, 2012; Gutmann and Hyvärinen, 2013). However, compared with the present study, these results were somewhat preliminary in terms of the number of reproduced experiments, compatibility with the experimental data, and faithfulness to the experimental protocol. Further, the size of the model and the variety of data were limited, and model variation was not considered in these studies, which leaves the scalability question unanswered. Finally, none of these studies formally analyzed the internal representation in detail and quantitatively related it to the neural response properties.

Although our model succeeded in capturing basic receptive field properties found in V2, it cannot exhibit certain kinds of reported complex nonlinear behavior as it is a simple half-rectified linear model taking V1 outputs. In particular, it did not explain the highly nonlinear interactions between local orientations at different positions reported by Anzai et al. (2007); the multiplicative behaviors reported by Ito and Goda (2011) and the "second-order kernels" reported by Schmid et al. (2014) are also unlikely to be reproducible. Although a relatively small population exhibits such nonlinearities according to these reports, this is still an important difference from V1 and might have significant implications in higher visual processing. Such nonlinearities might be related to nonlinear inference in sparse coding models (Olshausen and Field, 1996), but might also be related to other kinds of statistical modeling of natural images based on divisive normalization (Schwartz and Simoncelli, 2001) or feedback processing (Rao and Ballard, 1999; Hosoya, 2012), or to further highly nonlinear models yet to be developed.

Although we have focused on explaining the neural properties in V2 that were related to excitatory and inhibitory organization of local orientations, at least two other classes of tuning properties seem worth investigating. First, several properties related to artificial and naturalistic texture features have been studied in V2 (El-Shamayleh and Movshon, 2011; Freeman et al., 2013; Li et al., 2014). Since texture representations presumably involve not only orientation organization but also frequency and phase organization, detailed analysis of such representations learned from natural images and comparison to the known neural properties would be particularly interesting. Second, the responses of some V2 cells are modulated by stimulus parts that are quite far from their classic receptive fields; for example, surround suppression (Shushruth et al., 2009) and border ownership signals (Zhou et al., 2000). Since such surround effects are usually attributed to lateral or feedback connections, studying them from the viewpoint of natural image statistics might need more sophisticated models (but see Olshausen and Field, 1997; Schwartz and Simoncelli, 2001).

Finally, how far can we go beyond V2? Although the present study has demonstrated that a purely bottom-up learning model can explain well several neuronal properties in V2, it is not clear whether we can reach all the way to inferotemporal areas with this approach since these areas contain much more specialized representations, such as faces (Tsao et al., 2006), body parts (Pinsk et al., 2005), and scenes (Kornblith et al., 2013). To explain such high-level representations, some additional sources of information may have to be exploited, such as contextual information, attention, or "genetically" built-in information on evolutionary utility. While it is possible that only learning in the highest layers needs such mechanisms, it could also be that top-down influence sends such information down to V1, refining even the most basic representations—a question to be addressed in our future research.

References

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

Barlow HB (1961) Possible principles underlying the transformation of sensory messages. In: Sensory communication (Rosenblith WA, ed), pp 217–234. Cambridge, MA: MIT.

Berkes P, Orbán G, Lengyel M, Fiser J (2011) Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. Science 331:83–87. CrossRef Medline

Cadieu CF, Olshausen BA (2012) Learning intermediate-level representations of form and motion from natural movies. Neural Comput 24:827–866. CrossRef Medline

DeAngelis GC, Freeman RD, Ohzawa I (1994) Length and width tuning of neurons in the cat's primary visual cortex. J Neurophysiol 71:347–374. Medline

Deng J, Berg AC, Li K, Fei-Fei L (2010) What does classifying more than 10,000 image categories tell us? Comput Vis ECCV 6315:71–84. CrossRef

El-Shamayleh Y, Movshon JA (2011) Neuronal responses to texture-defined form in macaque visual area V2. J Neurosci 31:8543–8555. CrossRef Medline

Freeman J, Simoncelli EP (2011) Metamers of the ventral stream. Nat Neurosci 14:1195–1201. CrossRef Medline

Freeman J, Ziemba CM, Heeger DJ, Simoncelli EP, Movshon JA (2013) A functional and perceptual signature of the second visual area in primates. Nat Neurosci 16:974–981. CrossRef Medline

Gutmann MU, Hyvärinen A (2013) A three-layer model of natural image statistics. J Physiol Paris 107:369–398. CrossRef Medline

Hallum LE, Movshon JA (2014) Surround suppression supports second-order feature encoding by macaque V1 and V2 neurons. Vision Res 104:24–35. CrossRef Medline

Hosoya H (2012) Multinomial Bayesian learning for modeling classical and nonclassical receptive field properties. Neural Comput 24:2119–2150. CrossRef Medline

Hoyer PO, Hyvärinen A (2000) Independent component analysis applied to feature extraction from colour and stereo images. Network 11:191–210. CrossRef Medline

Hoyer PO, Hyvärinen A (2002) A multi-layer sparse coding network learns contour coding from natural images. Vision Res 42:1593–1605. CrossRef Medline

Hyvärinen A (2005) Estimation of non-normalized statistical models by score matching. J Mach Learn Res 6:695–709.

Hyvärinen A, Hoyer P (2000) Emergence of phase-and shift-invariant fea-

tures by decomposition of natural images into independent feature subspaces. Neural Comput 12:1705–1720. CrossRef Medline

Hyvärinen A, Hoyer PO (2001) A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. Vision Res 41:2413–2423. CrossRef Medline

Hyvärinen A, Karhunen J, Oja E (2001) Independent component analysis. New York: Wiley-Interscience.

Hyvärinen A, Gutmann M, Hoyer PO (2005) Statistical model of natural stimuli predicts edge-like pooling of spatial frequency channels in V2. BMC Neurosci 6:12. CrossRef Medline

Hyvärinen A, Hurri J, Hoyer PO (2009) Natural image statistics: a probabilistic approach to early computational vision. New York: Springer.

Ito M, Goda N (2011) Mechanisms underlying the representation of angles embedded within contour stimuli in area V2 of macaque monkeys. Eur J Neurosci 33:130–142. CrossRef Medline

Ito M, Komatsu H (2004) Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. J Neurosci 24:3313–3324. CrossRef Medline

Karklin Y, Lewicki MS (2003) Learning higher-order structures in natural images. Network 14:483–499. CrossRef Medline

Karklin Y, Lewicki MS (2009) Emergence of complex cell properties by learning to generalize in natural scenes. Nature 457:83–86. CrossRef Medline

Kornblith S, Cheng X, Ohayon S, Tsao DY (2013) A network for scene processing in the macaque temporal lobe. Neuron 79:766–781. CrossRef Medline

Köster U, Hyvärinen A (2010) A two-layer model of natural stimuli estimated with score matching. Neural Comput 22:2308–2333. CrossRef Medline

Lee H, Ekanadham C, Ng AY (2008) Sparse deep belief net model for visual area V2. In: Advances Neural Information Processing Systems, vol. 20 (Platt JC, Koller D, Singer Y, Roweis ST, eds.), pp 873–880.

Li G, Yao Z, Wang Z, Yuan N, Talebi V, Tan J, Wang Y, Zhou Y, Baker CL Jr (2014) Form-cue invariant second-order neuronal responses to contrast modulation in primate area V2. J Neurosci 34:12081–12092. CrossRef Medline

Nishimoto S, Ishida T, Ohzawa I (2006) Receptive field properties of neurons in the early visual cortex revealed by local spectral reverse correlation. J Neurosci 26:3269–3280. CrossRef Medline

Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381:607–609. CrossRef Medline

Olshausen BA, Field DJ (1997) Sparse coding with an overcomplete basis set: a strategy employed by V1? Vision Res 37:3311–3325. CrossRef Medline

Paninski L (2003) Convergence properties of three spike-triggered analysis techniques. Network 14:437–464. CrossRef Medline

Pinsk MA, DeSimone K, Moore T, Gross CG, Kastner S (2005) Representations of faces and body parts in macaque temporal cortex: a functional MRI study. Proc Natl Acad Sci U S A 102:6996–7001. CrossRef Medline

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79–87. CrossRef Medline

Schmid AM, Purpura KP, Victor JD (2014) Responses to orientation discontinuities in V1 and V2: physiological dissociations and functional implications. J Neurosci 34:3559–3578. CrossRef Medline

Schwartz O, Simoncelli EP (2001) Natural signal statistics and sensory gain control. Nat Neurosci 4:819–825. CrossRef Medline

Shushruth S, Ichida JM, Levitt JB, Angelucci A (2009) Comparison of spatial summation properties of neurons in macaque V1 and V2. J Neurophysiol 102:2069–2083. CrossRef Medline

Tao X, Zhang B, Smith EL 3rd, Nishimoto S, Ohzawa I, Chino YM (2012) Local sensitivity to stimulus orientation and spatial frequency within the receptive fields of neurons in visual area 2 of macaque monkeys. J Neurophysiol 107:1094–1110. CrossRef Medline

Tsao DY, Freiwald WA, Tootell RB, Livingstone MS (2006) A cortical region consisting entirely of face-selective cells. Science 311:670–674. CrossRef Medline

van Hateren JH, van der Schaaf A (1998) Independent component filters of natural images compared with simple cells in primary visual cortex. Proc Biol Sci 265:359–366. CrossRef Medline

Vinje WE, Gallant JL (2000) Sparse coding and decorrelation in primary visual cortex during natural vision. Science 287:1273–1276. CrossRef Medline

Willmore BD, Prenger RJ, Gallant JL (2010) Neural representation of natural images in visual area V2. J Neurosci 30:2102–2114. CrossRef Medline

Zhou H, Friedman HS, von der Heydt R (2000) Coding of border ownership in monkey visual cortex. J Neurosci 20:6594–6611. Medline