FOUNDATIONS OF VISUAL FORM SELECTIVITY FOR NEURONS IN MACAQUE V1 AND V2

A PREPRINT

Timothy D. Oleskiw
Center for Neural Science, New York University
Center for Computational Neuroscience, Flatiron Institute
oleskiw@nyu.edu

Justin D. Lieber
Center for Neural Science, New York University
justinlieber@nyu.edu

Eero P. Simoncelli
Center for Computational Neuroscience, Flatiron Institute
Center for Neural Science, New York University
eps2@nyu.edu

J. Anthony Movshon
Center for Neural Science, New York University
movshon@nyu.edu

March 4, 2024

ABSTRACT

We have measured the visually evoked activity of single neurons recorded in areas V1 and V2 of awake, fixating macaque monkeys, and captured their responses with a common computational model. We used a stimulus set composed of "droplets" of localized contrast, band-limited in orientation and spatial frequency; each brief stimulus contained a random superposition of droplets presented in and near the mapped receptive field. We accounted for neuronal responses with a 2-layer linear-nonlinear model, representing each receptive field by a combination of orientation- and scale-selective filters. We fit the data by jointly optimizing the model parameters to enforce sparsity and to prevent overfitting. We visualized and interpreted the fits in terms of an "afferent field" of nonlinearly combined inputs, dispersed in the 4 dimensions of space and spatial frequency. The resulting fits generally give a good account of the responses of neurons in both V1 and V2, capturing an average of 40% of the explainable variance in neuronal firing. Moreover, the resulting models predict neuronal responses to image families outside the test set, such as gratings of different orientations and spatial frequencies. Our results offer a common framework for understanding processing in the early visual cortex, and also demonstrate the ways in which the distributions of neuronal responses in V1 and V2 are similar but not identical.

1 Introduction

In the primate brain, complex visual patterns are processed by a cascade of computations in a set of visual areas collectively known as the ventral stream of visual cortex [9, 28, 47]. At the beginning of this cascade is primary visual cortex (V1), where neurons respond selectively to spatially-localized patterns of a specific orientation and scale [19]. Classically, these cells have been broadly split into two groups. Simple cells respond selectively to the local phase of an oriented pattern, and are well-described by a linear-nonlinear (LN) model: a single spatial filter followed by a rectification step[21, 30]. Complex cells respond independently of local phase, and are well-described by the squared sum of two spatial filters (the energy model) [1, 29]. In practice, many cells in V1 lie on a continuum between these two extremes [26, 34]. The full set of cells across the spectrum of V1 responses can be described by a unified model that spatially pools local populations of rectified linear filters [49]. These models, when combined with divisive normalization and contextual modulation by the receptive field surround, are state-of-the-art in predicting the responses of V1 neurons to natural images [5].

Extrastriate visual area V2, directly downstream of V1, has been implicated in the processing of visual patterns that are more complex than simple oriented lines. Like neurons in V1, many neurons in V2 respond selectively to spatially

localized patterns of specific orientations and spatial frequencies[22]. However, V2 neurons often respond with wider orientation and spatial frequency bandwidths than neurons in V1, and with higher contrast sensitivities, as would be expected if V2 neurons were pooling the responses of multiple V1 inputs [22]. Furthermore, lesions to V2 result in deficits to higher-order feature detection while preserving orientation perception [27], suggesting a specific, causal role for V2 in the perception of complex form. When probed with more complex stimuli, some V2 neurons respond selectively to curved contours [16] [2], object boundaries [54], and multipoint correlations in checkerboard textures [53]. We have recently shown that many neurons in V2 respond selectively to the higher-order statistics of "naturalistic" textures, while neurons in V1 do not [12].

Fitting models that can capture V2's selectivity for more complex visual features, i.e. form, may require a more sophisticated model structure than would be used in V1. It has long been recognized that a sufficient model of visual pattern recognition would require a multi-stage network of canonical computations [13, 14], with a first layer that approximates V1-like linear filters. These multi-stage models can explain perceptual phenomena, such as second-order texture boundary detection [4, 10, 50].

Models of this type have been fit to the responses of neurons in V2 [23, 36, 44, 51]. However, several factors make these fits difficult to interpret. First, many of these studies use images of natural scenes [23, 36, 51], which are highly correlated in the parameter space of these models. While stimulus bias can, in principle, be addressed by estimating the prior distribution over a parameter space, in practice, this requires many more presentation trials than can be collected in an awake experimental preparation. Second, these approaches have used rapid sequential stimulus presentations to estimate a neuron's receptive field in space and time [8, 31]. Given the dynamic selectivity of V1 neurons[25, 35], particularly over the extra-classical receptive field [7, 17, 38, 43, 52], it is unclear whether V2 response patterns measured using rapidly changing stimuli will extrapolate to more natural viewing conditions. Finally, interpreting model fits can be difficult when filters in intermediate layers do not cleanly map to earlier levels of neural processing [5].

In this study, we sought to overcome these challenges and to produce an interpretable account of neuronal receptive fields in macaque V1 and V2. We designed a stimulus set – *droplets* – to robustly activate neurons in V1 and V2 while containing only limited correlations in image space, allowing for unbiased modeling. We presented stimuli at an ethologically relevant cadence, similar to the typical primate inter-saccade interval [3]. To aid interpretability, we designed a model structure that fits neurons as a sparse combination of rectified filters that are selective for orientation and spatial frequency. Neurons in both V1 and V2 were well driven by the droplet stimuli and well fit by the model. Neurons in V1 were more tightly tuned in orientation and spatial frequency. In V2, but not V1, we found a population of neurons that pool broadly across different orientations and are selective for more complex image features.

2 Methods

2.1 Stimulus design

Visual stimuli were constructed by superimposing patches of orientation energy across the region of visual space encompassing a neuron's receptive field. In practice, spatial frequency elements were positioned on hexagonal latices of different spatial scales, comprised of three (coarse), five (mid-range), and seven (fine) elements to a side (Fig. 1a). These latices were scaled such that the central element of the coarse lattice occupied the estimated classical receptive field of the recorded unit (Fig. 1c). Each element was constructed from a sinusoidal grating subject to a raised cosine spatial envelope. Element orientation is randomly chosen from six equally-spaced orientations (30°) along the half-circle, including the cardinal horizontal and vertical. Similarly, each element's phase is randomly chosen from one of four quarter-cycle increments. Each element's spatial frequency was defined in units of cycles-per-element and thus varies across latices. However, each element could be of two spatial frequencies, either σ or 2σ, where σ is a free parameter chosen to agree with the unit's preferred spatial frequency from a set of previously generated stimuli families. The amplitude of an element was set to maximum luminance contrast, or zero, by randomly sampling a binomial random variable of probability chosen such that each lattice had an equal expected coverage of elements. In practice, we found cells to exhibit robust responses to a sparse sampling of elements (20%) across a lattice. The elements were spaced such that the raised cosine envelope produced an approximately uniform expected contrast across the stimulus. Elements were linearly summed in luminance and clipped to the dynamic range of the display.

2.2 Neuronal recordings

Neurophysiology data was collected from area V2 of two awake-and-fixating adult male rhesus macaques. Before experimentation, a custom headpost was surgically implanted for head stabilization using a standard design and methods described previously (Grey Matter Research. In a subsequent surgical procedure, a recording chamber was implanted

over the Lunate sulcus of the right hemisphere. Chamber placement was guided with structural magnetic resonance imaging (MRI) and visualization software (Brainsight, Rogue Research) to design a chamber with legs matched to the curvature of the monkey's skull above the lunate sulcus. All procedures complied with the National Institute of Health Guide for the Care and Use of Laboratory Animals, with the approval of the New York University Animal Welfare Committee.

We acclimated each monkey to his recording chair and experimental surroundings. After this initial period, he was head-restrained and rewarded for looking at the fixation target with dilute juice or water. Meanwhile, we used an infrared eye tracker (EyeLink 1000; SR Research) to monitor eye position at 1000 Hz via reflections of infrared light on the cornea and pupil. The monkey sat 57 cm from the display.

The monkey initiated a trial by fixating on a small white spot (diameter 0.1-0.2°), after which he was required to maintain fixation for a 200-500ms interval. A random droplet stimulus would appear for 200ms, followed by a 200ms inter-stimulus interval (Fig. 1c). The monkey was rewarded if he maintained fixation within 1-1.75° from the fixation point for the entire stimulus duration of 4-8 presentations. No stimuli were presented during the 300–600 ms in which the reward was delivered. If the monkey broke fixation prematurely, the trial was aborted, a timeout of 2000 ms occurred, and no reward was given.

Recordings were conducted by advancing a 6-10 MΩ impedance tungsten-epoxy microelectrode (FHC) through a 23 Gauge stainless steel guide that was stabilized with a customized 3D-printed grid insert affixed to the recording chamber. Both the grid and guide tube were held in contact with the dura during electrode penetration. We distinguished V2 from V1 based on depth from the cortical surface and changes in the receptive field location of recorded units. To obtain an unbiased sample of single units, we made extracellular recordings from every single unit with a spike waveform that rose sufficiently above background noise to be isolated. Data are reported from every unit for which we completed characterization. The receptive fields of most units were between 1° and 5° eccentricity, but our estimates of eccentricity and size were not sufficiently precise to include in analyses.

After isolating a single-unit spike waveform, receptive field location and size were estimated by hand. Tuning for orientation and spatial frequency is performed by presenting localized sinusoidal gratings at six orientations ($0 - 160^\circ$), five spatial frequencies ($.25 - 4$ cyc/°), and four equidistant phases, randomly interleaved at 200ms on/off intervals. The set of droplet stimuli was chosen to match the cell's preferred range of spatial frequencies.

Neural response to each stimulus was determined from the mean evoked activity, estimated from each unit's peristimulus time histogram (PSTH) computed for stimulus-present and stimulus-absent (baseline) trials (Fig. 1b). The response is

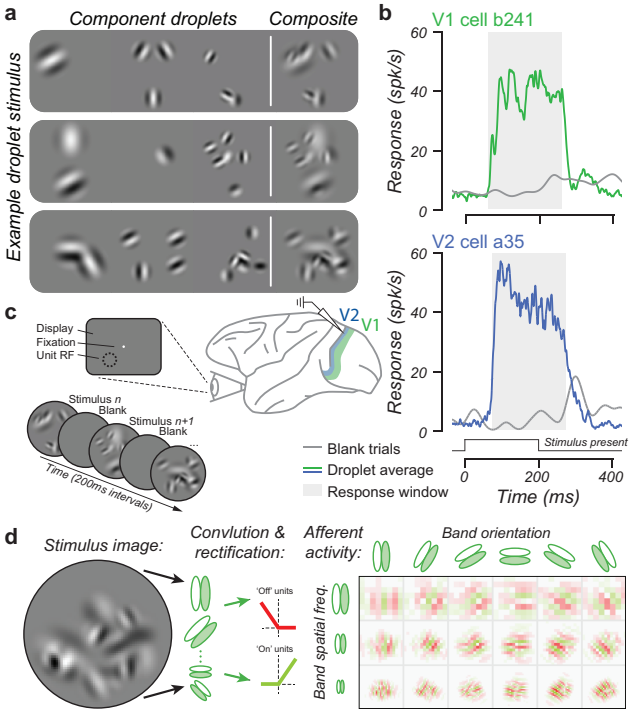

Figure 1: Stimulus design and recording paradigm. (a) Droplet stimuli are constructed from multi-scale patches of gratings tiling space on hexagonal grids. (b) Example V1 and V2 neuron responses to an interleaved presentation (c) of droplet stimuli. (d) By localizing contrast in space, orientation, and scale, this stimulus class sparsely activates a V1-like afferent space

determined by calculating the expected number of spikes within a time window beginning at the unit's response latency and lasting for the stimulus duration (200 ms).

2.3 An afferent model of visual computation

We consider an image-computable two-layer linear-nonlinear network resembling a neural unit that receives input from a population of rectified linear filters tuned in orientation and scale. For a set of K stimuli, We denote the observed response to the k^{th} stimulus S_k as r_k. The prediction p of the model is defined by

$$p\left(S_k\right) = f_\theta\left(g\left(S_k\right) \cdot w\right) \tag{1}$$

where f_θ denotes the output nonlinearity parameterized by θ, and g denotes the first-layer activation of stimulus S_k to be pooled via linear combination with connection weights w. We choose f_θ to be a piece-wise linear two-sided rectifier,

$$f_\theta(x) = \beta + \alpha^+ \max(0, x) + \alpha^- \max(0, -x), \tag{2}$$

with parameters $\theta = \{\alpha^+, \alpha^-, \beta\}$ encoding the positive gain, negative gain, and offset, respectively.

To model first layer responses g we utilize the steerable pyramid transform [40], spatially convolving a bank of linear filters tuned in orientation and scale. This decomposes an image into multiple spectral bands, each localized to a preferred orientation and spatial frequency, thereby mimicking a population of V1-like oriented units that tile both spatial and spectral domains of the image (Fig. 1d). We denote the pyramid coefficients for M spectral bands of the stimulus S_k as

$$g\left(S_k\right) = X = \{x_1, \ldots, x_M\}. \tag{3}$$

Thus, X is a collection of steerable pyramid coefficient vectors. To mimic V1 simple cell activation, coefficient vectors for each band are positive and negative half-wave rectified, *i.e.*,

$$x_i = \{x_i^+, x_i^-\} = \{\max(0, x_i), \max(0, -x_i)\}, \tag{4}$$

resulting in $2M$ coefficient vectors, written $X = \left\{x_1^+, \ldots, x_M^+, x_1^-, \ldots, x_M^-\right\}$. Intuitively, X represents the activity from a population of V1-like simple cells, each tuned in orientation and scale, that tile horizontal and vertical space. Next, to pool first-layer activity, a connection weight vector w, operating across each spatial dimension and rectified band, computes the linear combination of transformed pyramid coefficients. Specifically, for every band $m \in M$, the inner product is taken between a pair of filters, w_m^+ and w_m^-, and the corresponding rectified spectral band coefficients, x_m^+ and x_m^-. The activity of the second layer unit is therefore computed by summing this weighted response over all bands, passing the resulting generator signal through the output nonlinearity f_θ. Thus, the output response of a model unit to stimulus S_k is governed by

$$p(S_k; w, \theta) = f_\theta\left(\sum_{m \in M} x_m^+ \cdot w_m^+ + x_m^- \cdot w_m^-\right), \tag{5}$$

parameterized by rectified connection weights $w = \{w^+, w^-\}$ and output nonlinearity terms $\theta = \{\alpha^+, \alpha^-, \beta\}$.

2.4 Model training and regularization

To optimize the afferent model, we seek to find the connection weights w and nonlinear parameters θ to minimize the squared error between model predictions $\hat{r}_{w,\theta}$ and observed responses r for all K stimulus-response pairs:

$$\|r - \hat{r}_{w,\theta}\|_2 = \left(\sum_{k=1}^{K} \left(r_k - p(S_k; w, \theta)\right)^2\right)^{\frac{1}{2}} \tag{6}$$

Since each model can have thousands of connection weight parameters, much greater than the number of unique stimuli presented to a neuron, this optimization problem is undetermined and susceptible to over-fitting. To address this, a sparsity L_1 regularization term is introduced to penalize the magnitude of model connection weights, *i.e.*, $\lambda\|w\|$ for penalty magnitude λ. Since coefficients at every spatial location and spectral band are matched across rectification channels, a group regularization constraint is employed to combine weights across channels. Here, the *magnitude* of filter weights is to be penalized, but the *relative contribution* of positive and negative rectified channels is unconstrained. Said another way, the model may freely choose the scaled rectification of each V1-like unit of the first layer. The connection weight magnitude for regularization is defined by

$$w_{i,m}^* = \sqrt{\left(w_{i,m}^+\right)^2 + \left(w_{i,m}^-\right)^2} \tag{7}$$

for the i^{th} pyramid coefficient of band $m \in M$. However, the magnitude of connection weights can be made arbitrarily small while maintaining prediction accuracy by increasing the gain parameters α_+ and α_- of the output nonlinearity. To remove this degree of freedom from the solution space, we normalize connection weight sparsity by the total connection energy. To fit a model neuron, we optimize for w and θ, minimizing

$$\|r - \hat{r}_{w,\theta}\|_2 + \lambda \frac{\|w^*\|}{\|w^*\|_2}. \tag{8}$$

Under this formulation, the regularization strength λ constrains the dimensionality of connection weights w, forcing $|w| \to 0$ as $\lambda \to \infty$. It is important to note, however, that each weight is regularized independently, disregarding relationships between spectrally- or spatially-neighboring coefficients. We therefore wish to bias the regularization in such a way as to prefer afferent maps that are localized in the space of afferents. To achieve this, we approximate the spectral and spatial dispersion of coefficient magnitudes from the marginal variances of weights. First, spectral bands are referenced in octaves, and spatial coordinates are scaled such that an octave represents twice the estimated receptive field diameter to equate the units of these variances. Regularization enforcing afferent sparsity and locality is therefore computed via

$$\lambda \left[\frac{\|w^*\|}{\|w^*\|_2} + \gamma \left(\sqrt{\sigma_s^2 + \sigma_o^2} + \sqrt{\sigma_h^2 + \sigma_v^2} \right) \right], \tag{9}$$

where afferent dispersion is denoted σ_s^2, σ_o^2 for scale and orientation, and σ_h^2, σ_v^2 for horizontal and vertical space, respectively. Note that locality bias parameter γ is empirically chosen to penalize afferent dispersion at approximately 1% the magnitude of coefficient sparsity. This regularization formulation had the qualitative effect of biasing optimized models toward sparse and local afferent maps without significantly impacting convergence or validation performance.

Given that there are many more connection weight parameters than stimulus-response data points, we can choose λ to constrain the solution space and prevent over-fitting. We then select an optimal λ using a cross-validation method to guarantee that constrained solutions generalize across our dataset. Here, stimulus-response pairs are randomly assigned to one of ten equally sized partitions. For each partition, a model is trained from the remaining stimulus-response pairs, then used to predict the held-out responses. This process is repeated for multiple regularization strengths λ. Training (fitting) and testing errors are computed from the sum of squared error between observed responses and model predictions, normalized by the variance of observed responses. For each neuron, we choose λ' for each neuron to maximize the withheld (testing) variance explained, limiting the dimensionality of w while ensuring a robust fitted model.

3 Results

3.1 V1 and V2 neurons are well explained by sparse & localized combinations of afferent activity

We quantify the ability of our two-layer linear-nonlinear model to explain V1 and V2 activity by plotting the mean variance explained (V.E.) across training and testing partitions. Fig. 2a,c depicts training performance as a function of testing performance, demonstrating both measures to be highly correlated across our population, inconsistent with model over-fitting. A fraction of the cells in our population were poorly fit by the model (V1:$n = 12$ of 69, V2: $n = 20$ of 120) exhibited a low training (< 0.1) and testing (< 0) performance. Upon inspection, these units had too few stimulus repetitions or low evoked activity to allow for effective model fitting. Moreover, some of these units were most active at the offset of a stimulus, which was not captured by the response activity window and thus was not accounted for in our analysis. To simplify our analyses and comparison of units, we consider a subset of well-fit V1 and V2 units (testing V.E > 0.1). The mean training/testing performance by our model to this population is 0.36/0.27 for V1 ($n = 32$), and 0.38/0.26 for V2 ($n = 56$), respectively.

While the reported variance explained appears modest, it is masked by the intrinsic variability of neurons and the limited number of repeated trials of a given stimulus. To measure the amount of *explainable variance*, *i.e.* the predictive power of a model up to independent neuronal variability, we first estimate the noise ceiling of each recording. Here, for each neuronal dataset, we identify stimuli with three or more repeated trials, computing the mean and variance of response for each stimulus. To model trial-to-trial variability, we assume scaled-Poisson noise and find the optimal scalar mapping between a neuron's response mean and variance for each repeated stimulus. Then, we simulate a *synthetic* recording session, sampling firing rates under the fitted noise model for every recorded trial. By averaging across repeated stimulus trials, we effectively sample neuronal activity from an identical set of trials as was recorded. This process is boot-strap repeated and compared against observed responses to estimate the distribution of the expected recording error. In essence, this procedure estimates expected variability having repeated an identical recording session, serving as an upper bound for the response variance that is resolvable by any model. By normalizing the testing variance explained by this noise ceiling, we plot in Fig. 2a,c this explainable variance. Many neurons generalize well to capture

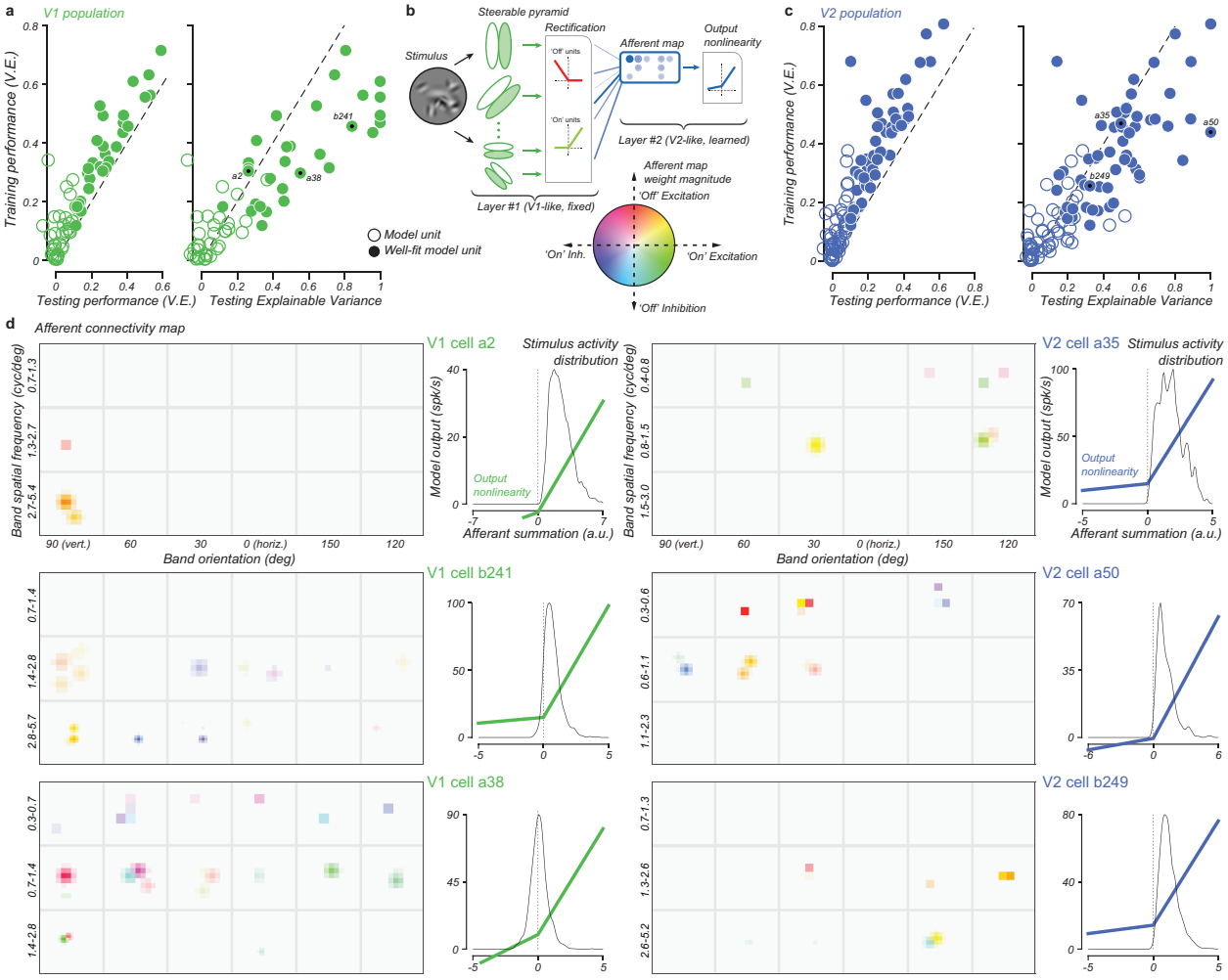

Figure 2: Model Performance and example unit fits. Training and testing performance of models fit to (a) V1 and (c) V2 neurons in our population. (b) Each model optimized the connection weights projecting from a V1-like afferent representation. (c) Model fits are visualized by an afferent map and output nonlinearity for three example V1 and V2 cells. The hue and saturation of afferent weights convey the connection strength of the on and off rectification channels, simultaneously.

a considerable fraction of the stimulus-response variance expected to be explained by a model. Of units that were sufficiently driven by our stimulus for the model to converge, the average explainable variance captured by the model was 40% (V1: 43%, V2: 38%). For well-fit units considered further for analysis, the average explainable variance was 51% (V1: 56%, V2: 48%).

To visualize optimized model parameters, we depict in Fig. 2d a representation of the connection weights w and output nonlinearity f_θ of example V1 and V2 units. Connection weights are grouped into spectral bands of orientation (columns) and spatial frequency (rows). Within each band, weights are arrayed in their natural horizontal and vertical spatial position, *i.e.* each band represents the same region of visual space. We then superimpose the two independent rectified channels for each spectral band, using saturation to denote absolute connection magnitude and hue to encode the relative contribution of each channel. We denote this representation of projection weights, pooling contrast across spectral and spatial domains, the *afferent map*.

3.2 Linear and separability of afferent maps

The afferent map of connection weights, representing a linear combination of basis elements tuned in orientation and spatial frequency that tile visual space, is segregated into positively and negatively rectified terms. However, given that afferents are linearly combined, we can change the basis to organize afferent activity along arbitrary axes in the

afferent space. One natural basis for consideration is the linear and energy space, constructed from half-wave rectified coefficient vectors x^+ and x^- via

$$\begin{cases} x^l = \frac{1}{\sqrt{2}}\left(x^+ - x^-\right) \\ x^e = \frac{1}{\sqrt{2}}\left(x^+ + x^-\right) \end{cases} . \tag{10}$$

By similarly transforming the optimized connection weights w to yield w^l and w^e, we can explore how each model pools linear and nonlinear stimulus features.

Specifically, we determine the fraction of response variance explained by the afferent map is attributed to the linear and energy components in isolation. To do this, we project stimulus activity $g\left(S_k\right)$ for each $k \in K$ against w^l and w^e, computing the squared correlation coefficient (Pearson's r^2) between recorded neuronal activity. The linear and energy component map r^2 is normalized by the total afferent map r^2, up to the output nonlinearity, independently for each model. This fraction of variance explained by each model's linear and energy map components is given as Fig. 3a. Note, the fraction variance explained by the linear and energy components is *approximately* equal to 1 due to the output nonlinearity and coefficients shared between components. We define a linearity index by the difference of variance explained from the linear and energy component maps.

We next assess to what extent the connection weight maps are separable along the cardinal dimensions of afferent tuning. To achieve this, we approximate each afferent map with the product of three functions computed from the marginal of connectivity in 2D-space, orientation, and scale. We define a separability ratio of each model unit as the fraction of variance explained by the separable afferent map relative to that of the original map. For V1 and V2 model units, we find in Fig. 3b a significant relationship between afferent separability and linearity, with linear units tending to be inseparable along afferent tuning dimensions. This relationship appears most significant in V1, with no well-fit V1 model units being nonlinear and inseparable or separable and linear.

3.3 Visualizing model selectivity: the afferent field

As our analysis of the afferent map shows V2 model units to be well-characterized by their constituent linear or energy components, we seek a method to interpret the spatial and spectral organization of each component beyond the afferent map representation illustrated in Fig. 4a. Since the steerable pyramid basis is a linear transformation of pixels, the linear map w^l of pyramid coefficients can be inverted to visualize its contribution to model response in image coordinates. We denote the visualization of this component to be the model's linear receptive field, analogous to that achieved via reverse correlation to depict V1 simple cells. A solid circle is overlaid to depict the hand-mapped receptive field location and the field cropped to depict the spatial extent of the droplet stimuli.

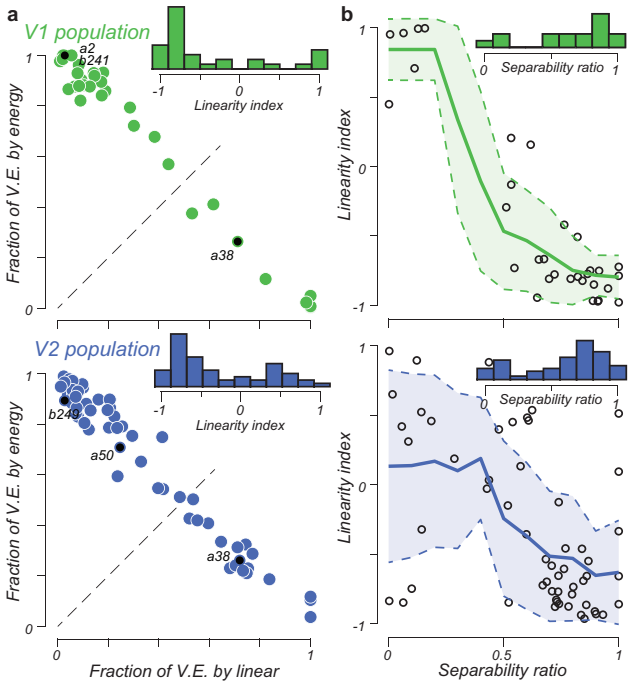

Figure 3: Model unit linearity and separability. (a) Units across the V1 and V2 populations span a continuum of selectivity, explained by a linear or nonlinear (energy) component of the afferent connectivity. (b) Linearity of a model unit's afferent map predicts the separability of afferents along the cardinal dimensions of orientation, scale, and space.

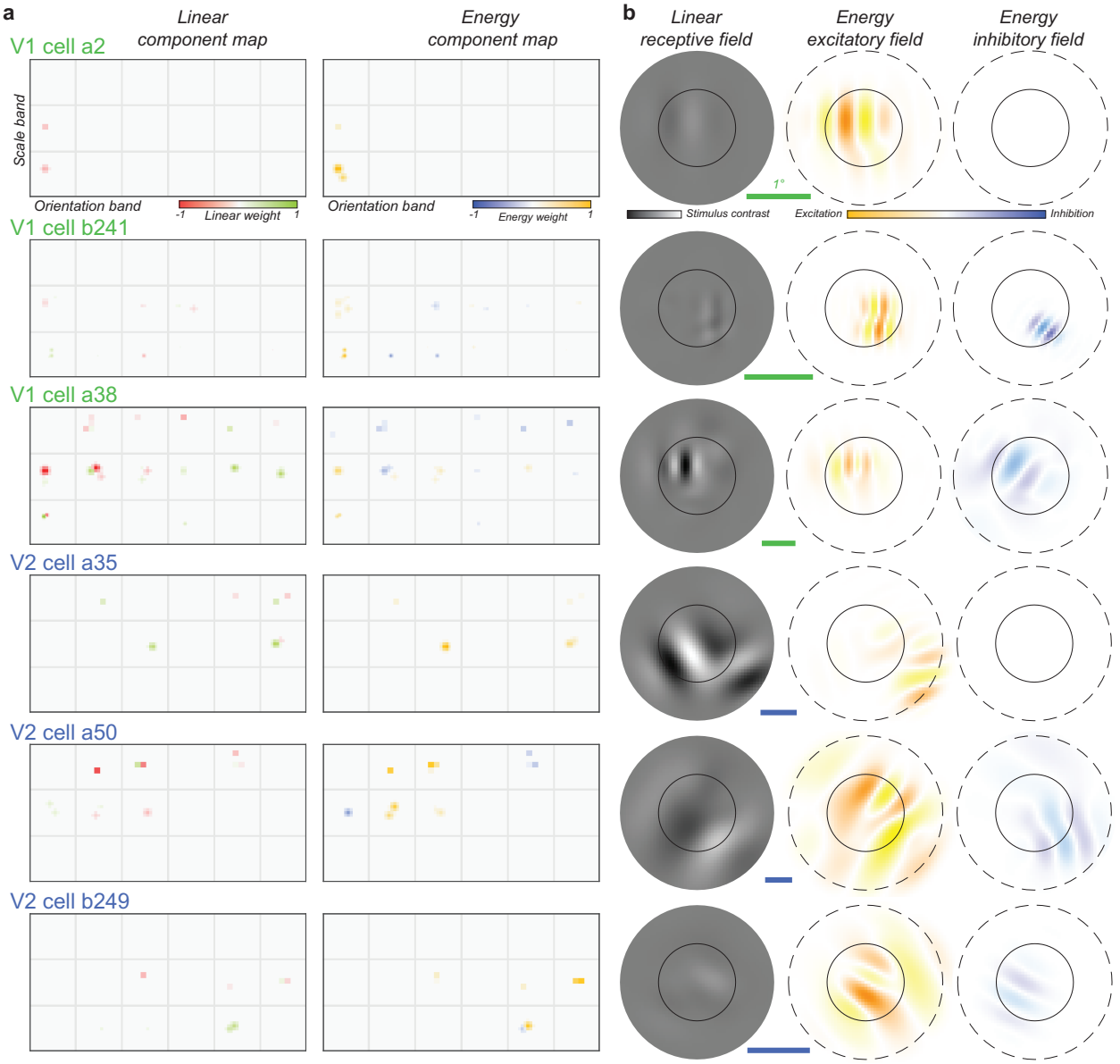

Figure 4: (a) Linear and energy components of example V1 and V2 afferent maps (see Fig. 2d). (b) Afferent field visualization of each unit's selectivity. Circles overlaying the linear and energy fields denote the cell's receptive field diameter as estimated from hand-mapping (solid) and droplet stimulus extent (dashed). Scale bars denote 1 degree of visual angle.

Interpretation of the energy component weights, however, cannot be achieved with the same visualization technique as multiple pixel representations of spectral energy exist up to the sign of all non-zero w^e terms. Instead, the spectral and spatial arrangement of w^e is depicted by producing a separate receptive field image for excitatory $(w^e > 0)$ and inhibitory $(w^e < 0)$ terms. A characteristic image is chosen that limits destructive interference due to the phase (sign) of pyramid coefficients. The sinusoidal image components are depicted with warm and cool hues to convey phase-invariant spectral power of the excitatory and inhibitory energy fields, respectively. Finally, the contrast of linear and energy receptive fields are scaled to convey their relative contribution to the model's response as determined from the variance explained by w^l and w^e, respectively, preserving the relative difference in magnitude between excitatory and inhibitory fields. Together, this afferent field, depicted in Fig. 4b, yields a spatial representation of a model unit's linear and energy components that comprise its selectivity.

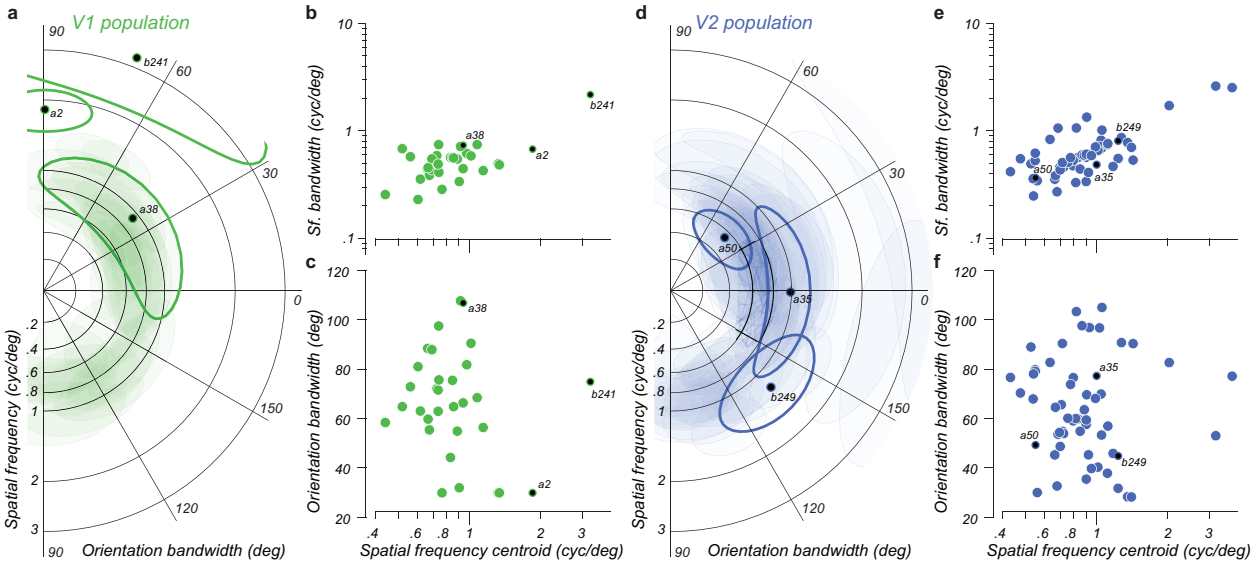

Figure 5: Spectral bandwidth of afferent connection weights from well-fit units from our population of (a) V1 and (d) V2 neurons. The bandwidth of (b,e) spatial frequency (c,f) and orientation is plotted as a function of the afferent map's spatial frequency centroid.

3.4 Spectral bandwidth of model afferents

To localize afferent connectivity, we compute the mean and variance of model weights in both orientation and spatial frequency. In Fig. 5, we plot spectral bandwidth as a function of preferred spatial frequency for well-fit units from our V1 and V2 populations. Fig. 5a,d shows our population to uniformly cover preferred orientations, with a distribution of spatial frequency coverage expected at the eccentricity of our recordings. As expected, we find a strong relationship between the bandwidth of spatial frequency connections and their mean (Fig. 5b,e). Conversely, no relationship is found between the mean of spatial frequency afferents and the bandwidth of orientation tuning (Fig. 5c,f).

An inspection between our V1 and V2 populations reveals no apparent qualitative difference between the spectral bandwidth of well-fit model afferents, with a similar distribution of units being more circular (*e.g.*, a2 and b249) or more elongated (*e.g.*, b241 and a35). This analysis, however, this analysis considers only two of the four dimensions of connection weights and ignores the potential covariance structure of the afferent map. We now assess the covariance structure of the afferent connectivity from well-fit model units.

3.5 Dimensionality of the afferent field

To assess the structure of afferent weight magnitudes in the four dimensions of selectivity, we compute the expected covariance between each pair of tuning dimensions. This calculation is complicated by the differing coordinates across dimensions, with only horizontal and vertical space being comparable. Within the limits imposed by its periodic nature, the orientation dimension can be represented in octaves, allowing a natural comparison to its spectral pair, spatial frequency (scale). To factor out the extent of each receptive field in visual space, we normalize the spatial covariance of each model to have unit norm. This representation is unique up to an arbitrary scaling of the spectral and spatial dimensions. We choose this scaling empirically from our population data to equate the mean spectral and mean spatial covariance across all well-fit model units from V1 and V2.

Having established a common coordinate frame to examine the structure of the afferent weight envelope, we next consider the eigenstructure of the unified covariance matrix calculated from each model unit. Qualitative inspection of the eigenvalue spectrum revealed units existed on a continuum between being were nearly spherical in four dimensions (*i.e.*, all eigenvalues of similar magnitude), and being elongated along in one dimension (*i.e.*, dominated by a single principal eigenvalue). To quantify this observation, we define the *dimensionality* of each unit in afferent space to be the (linearly-interpolated) number of eigenvectors required to capture 75% of coefficient weight variance. Fig. 6 depicts this dimensionality across visual areas, plotting the standard deviation of each afferent field in both scale and orientation (Fig. 6a,b). Interestingly, while the distribution of afferent dimensionality has a similar central tendency between V1 and V2 (Fig. 6c,e), a notable difference exists at the tails of each distribution: V1 has a higher frequency of units of

high dimensionality, *i.e.*, more spherical in afferent space, while V2 has more units that are more *prolate*, or elongated along a single afferent axis (Fig. 6d).

To investigate model unit prolation we examine the alignment of the principle axis for units that are highly elliptical in afferent space (*i.e.*, below a dimensionality of 1.4). We find prolate V1 units more aligned along spatial dimensions, having significant afferent weight covariance in space; *e.g.* unit a2 which has an elongated receptive field in the direction of its orientation tuning. In contrast, prolate V2 units have alignments spanning the orientation dimension, *e.g.* unit a35 which exhibits an afferent map highly correlated in space and orientation that is clearly evident in the linear component of its afferent field (Fig. 6g).

4 Discussion

Many different model structures have been used to fit the activity of neurons in V1. Early attempts used unbiased strategies to estimate the receptive field structure of V1 neurons, such as spike-triggered averaging [21] or spike-triggered covariance [11, 37]. These estimates suggested a simple, shared structure for most V1 receptive fields: most neurons could be described as the rectified sum of a set of self-similar, spatially shifted filters. A model that takes these assumptions into account uses fewer parameters than earlier methods, and fits V1 receptive fields more effectively and efficiently than previous methods [49].

The structure of receptive fields in V2 is not yet well established. However, we found it logical that V2 receptive fields should be parsimoniously described as a combination of V1-like inputs. Consistent with the model proposed by Vintch *et al.* [49], we find that many cells in V1 and V2 are well described as a local spatial pooling of self-similar, rectified filters. However, unlike the prior model, our model also has the flexibility to uncover a population of neurons in V2 that pool broadly across filter orientations. This is consistent with previous results showing that neurons in V2 were more often fit with combinations of orientations than neurons in V1 [23].

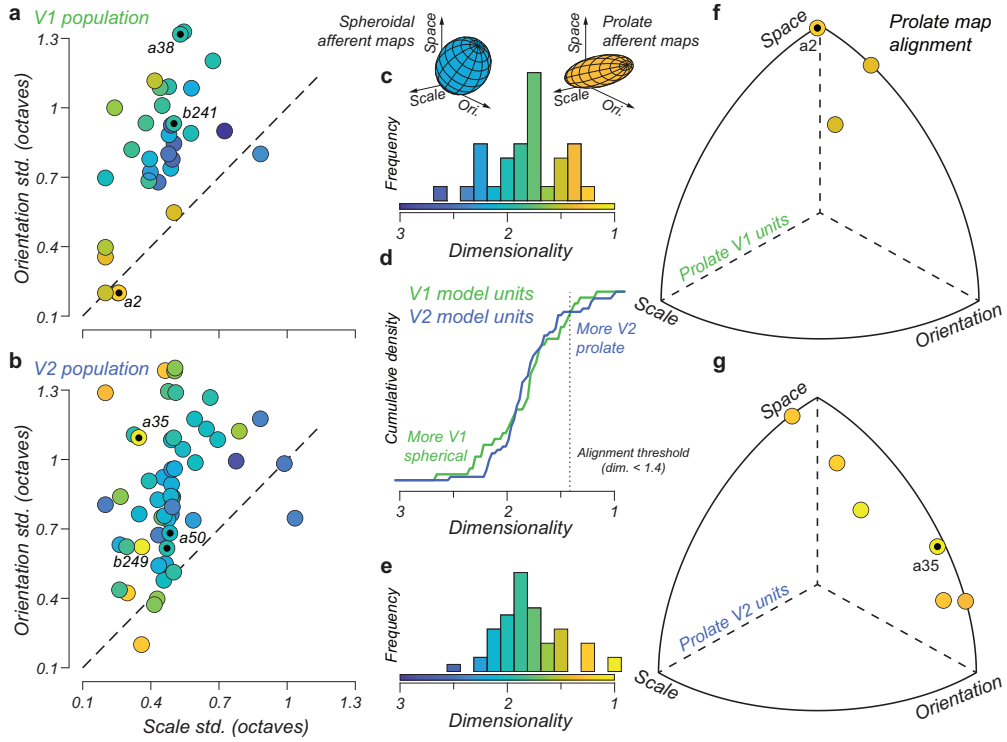

Figure 6: Covariance structure of model afferent weights. The standard deviation of afferent magnitudes in orientation and scale for units across our (a) V1 and (b) V2 populations, with color denoting (e,c) the dimensionality of each afferent field. (d) The cumulative distribution of afferent field dimensionality reveals that the most spherical units tend to be from V1, and most prolate tend to be from V2. (f,g) We visualize the alignment of prolate model units, defined as having a dimensionality below 1.4, as the orthant vector from the principal axis of afferent weight magnitude covariance.

Modeling populations of orientation- and scale-selective units for image representation has demonstrated that V1 neurons form an efficient code of natural visual scenes [32, 33, 39, 41, 48]. Efficient coding has also been used to predict the structure of neurons in V2 [6, 15, 18, 20].

We find many neurons in V2 that appear V1-like when probed with our stimuli. However, our stimuli were optimized to map the form-processing properties of these neurons. Some neurons in V2 display novel selectivities not observed in V1, such size-invariant chromatic selectivity [43], and relative disparity selectivity in three-dimensional scenes [45]. Anatomical studies illustrating the partition of V2 into different stripe compartments based on staining for the respiratory chain enzyme cytochrome oxidase [42], which suggests that there might be different functional clusters with distinct selectivities [23, 24, 46]. It may be that the "V1-like" neurons we observe in V2 are simply selective for features other than form.

Alternatively, V2 may just contain a large population of "V1-like" cells, and these complex selectivities might cluster within the same subpopulation of cells. In this case, neurons of V2 could be described as spanning a functional continuum from simpler to more complicated response properties, analogously to how V1 contains both "simple" and "complex" response properties. Future studies mapping multiple functional properties of these neurons be able to shed light on this question.

5 Acknowledgements

We are grateful to Kaitlyn Holman, Rui Pacheco, and Sullivan Bacerdo for their assistance. Manu Raghavan and Najib Majaj participated in some of the experiments and helped with surgery, hardware, and software. This work was supported by grants from the National Institutes of Health (EY022428) and the Simons Foundation (543019) to J.A.M. and E.P.S. J.D.L was supported in part by a Leon Levy Fellowship in Neuroscience.

References

[1] Adelson, E. H. and Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2(2):284.

[2] Anzai, A., Peng, X., and Van Essen, D. C. (2007). Neurons in monkey visual area V2 encode combinations of orientations. *Nature neuroscience*, 10(10):1313–21.

[3] Berg, D. J., Boehnke, S. E., Marino, R. A., Munoz, D. P., and Itti, L. (2009). Free viewing of dynamic stimuli by humans and monkeys. *Journal of Vision*, 9(5):19–19.

[4] Bergen, J. R. and Adelson, E. H. (1988). Early vision and texture perception. *Nature*, 333(6171):363–364.

[5] Burg, M. F., Cadena, S. A., Denfield, G. H., Walker, E. Y., Tolias, A. S., Bethge, M., and Ecker, A. S. (2021). Learning divisive normalization in primary visual cortex. *PLOS Computational Biology*, 17(6):e1009028.

[6] Cadieu, C. F. and Olshausen, B. A. (2012). Learning Intermediate-Level Representations of Form and Motion from Natural Movies. *Neural Computation*, 24(4):827–866.

[7] Cavanaugh, J. R., Bair, W., and Movshon, J. A. (2002). Nature and Interaction of Signals From the Receptive Field Center and Surround in Macaque V1 Neurons. *Journal of Neurophysiology*, 88(5):2530–2546.

[8] David, S. V., Vinje, W. E., and Gallant, J. L. (2004). Natural Stimulus Statistics Alter the Receptive Field Structure of V1 Neurons. *Journal of Neuroscience*, 24(31):6991–7006.

[9] DiCarlo, J., Zoccolan, D., and Rust, N. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, 73(3):415–434.

[10] DiMattina, C. and Baker, C. L. (2019). Modeling second-order boundary perception: A machine learning approach. *PLOS Computational Biology*, 15(3):e1006829.

[11] Felsen, G. and Dan, Y. (2005). A natural approach to studying vision. *Nature Neuroscience*, 8(12):1643–1646.

[12] Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., and Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, 16(7):974–981.

[13] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202.

[14] Granlund, G. H. (1978). In search of a general picture processing operator. *Computer Graphics and Image Processing*, 8(2):155–173.

[15] Gutmann, M. U. and Hyvärinen, A. (2013). A three-layer model of natural image statistics. *Journal of Physiology-Paris*, 107(5):369–398.

[16] Hegdé, J. and Van Essen, D. C. (2007). A comparative study of shape representation in macaque visual areas V2 and V4. *Cerebral Cortex*, 17(5):1100–1116.

[17] Henry, C. A., Jazayeri, M., Shapley, R. M., and Hawken, M. J. (2020). Distinct spatiotemporal mechanisms underlie extra-classical receptive field modulation in macaque V1 microcircuits. *eLife*, 9.

[18] Hosoya, H. and Hyvarinen, A. (2015). A Hierarchical Statistical Model of Natural Images Explains Tuning Properties in V2. *Journal of Neuroscience*, 35(29):10412–10428.

[19] Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–43.

[20] Hyvärinen, A., Gutmann, M., and Hoyer, P. O. (2005). Statistical model of natural stimuli predicts edge-like pooling of spatial frequency channels in V2. *BMC Neuroscience*, 6(1):12.

[21] Jones, J. P. and Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1187–1211.

[22] Levitt, J. B., Kiper, D. C., and Movshon, J. A. (1994). Receptive fields and functional architecture of macaque V2. *Journal of Neurophysiology*, 71(6):2517–2542.

[23] Liu, L., She, L., Chen, M., Liu, T., Lu, H. D., Dan, Y., and Poo, M.-m. (2016). Spatial structure of neuronal receptive field in awake monkey secondary visual cortex (V2). *Proceedings of the National Academy of Sciences*, 113(7):1913–1918.

[24] Livingstone, M. S. and Hubel, D. H. (1987). Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *J Neurosci*, 7(11):3416–68.

[25] Mazer, J. A., Vinje, W. E., McDermott, J., Schiller, P. H., and Gallant, J. L. (2002). Spatial frequency and orientation tuning dynamics in area V1. *Proceedings of the National Academy of Sciences*, 99(3):1645–1650.

[26] Mechler, F. and Ringach, D. L. (2002). On the classification of simple and complex cells. *Vision Research*, 42(8):1017–1033.

[27] Merigan, W., Nealey, T., and Maunsell, J. (1993). Visual effects of lesions of cortical area V2 in macaques. *The Journal of Neuroscience*, 13(7):3180–3191.

[28] Mishkin, M. and Ungerleider, L. G. (1982). Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behav Brain Res*, 6(1):57–77.

[29] Movshon, J. A., Thompson, I. D., and Tolhurst, D. J. (1978a). Receptive field organization of complex cells in the cat's striate cortex. *The Journal of Physiology*, 283(1):79–99.

[30] Movshon, J. A., Thompson, I. D., and Tolhurst, D. J. (1978b). Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *J Physiol*, 283:53–77.

[31] Nandy, A. S., Sharpee, T. O., Reynolds, J. H., and Mitchell, J. F. (2013). The Fine Structure of Shape Tuning in Area V4. *Neuron*, 78(6):1102–1115.

[32] Olshausen, B. A. and Field, D. J. (1996). Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7(2):333–339.

[33] Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research*, 37(23):3311–25.

[34] Priebe, N. J., Mechler, F., Carandini, M., and Ferster, D. (2004). The contribution of spike threshold to the dichotomy of cortical simple and complex cells. *Nature Neuroscience*, 7(10):1113–1122.

[35] Ringach, D. L., Hawken, M. J., and Shapley, R. (1997). Dynamics of orientation tuning in macaque primary visual cortex. *Nature*, 387(6630):281–284.

[36] Rowekamp, R. J. and Sharpee, T. O. (2017). Cross-orientation suppression in visual area V2. *Nature Communications*, 8(1):15739.

[37] Rust, N. C., Schwartz, O., Movshon, J. A., and Simoncelli, E. P. (2005). Spatiotemporal Elements of Macaque V1 Receptive Fields. *Neuron*, 46(6):945–956.

[38] Sceniak, M. P., Ringach, D. L., Hawken, M. J., and Shapley, R. (1999). Contrast's effect on spatial summation by macaque V1 neurons. *Nature Neuroscience*, 2(8):733–739.

[39] Schwartz, O. and Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–825.

[40] Simoncelli, E. P. and Freeman, W. T. (1995). Steerable pyramid: a flexible architecture for multi-scale derivative computation. *IEEE International Conference on Image Processing*, 3:444–447.

[41] Simoncelli, E. P. and Olshausen, B. A. (2001). Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience*, 24(1):1193–1216.

[42] Sincich, L. C. and Horton, J. C. (2005). THE CIRCUITRY OF V1 AND V2: Integration of Color, Form, and Motion. *Annual Review of Neuroscience*, 28(1):303–326.

[43] Solomon, S. G., Peirce, J. W., and Lennie, P. (2004). The Impact of Suppressive Surrounds on Chromatic Properties of Cortical Neurons. *Journal of Neuroscience*, 24(1):148–160.

[44] Tao, X., Zhang, B., Smith, E. L., Nishimoto, S., Ohzawa, I., and Chino, Y. M. (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(4):1094–1110.

[45] Thomas, O. M., Cumming, B. G., and Parker, A. J. (2002). A specialization for relative disparity in V2. *Nature Neuroscience*, 5(5):472–478.

[46] Ts'o, D. Y., Roe, A. W., and Gilbert, C. D. (2001). A hierarchy of the functional organization for color, form and disparity in primate visual area V2. *Vision Research*, 41(10-11):1333–1349.

[47] Ungerleider, L. G. and Mishkin, M. (1982). Analysis of visual behavior. In *Analysis of Visual Behaviour*, chapter Two cortic, pages 549–586. MIT Press.

[48] van Hateren, J. H. and Ruderman, D. L. (1998). Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1412):2315–2320.

[49] Vintch, B., Movshon, J. A., and Simoncelli, E. P. (2015). A Convolutional Subunit Model for Neuronal Responses in Macaque V1. *The Journal of Neuroscience*, 35(44):14829–14841.

[50] Voorhees, H. and Poggio, T. (1988). Computing texture boundaries from images. *Nature*, 333(6171):364–367.

[51] Willmore, B. D. B., Prenger, R. J., and Gallant, J. L. (2010). Neural representation of natural images in visual area V2. *J Neurosci*, 30(6):2102–14.

[52] Xing, D., Shapley, R. M., Hawken, M. J., and Ringach, D. L. (2005). Effect of Stimulus Size on the Dynamics of Orientation Selectivity in Macaque V1. *Journal of Neurophysiology*, 94(1):799–812.

[53] Yu, Y., Schmid, A. M., and Victor, J. D. (2015). Visual processing of informative multipoint correlations arises primarily in V2. *eLife*, 4:1–13.

[54] Zhou, H., Friedman, H. S., and von der Heydt, R. (2000). Coding of Border Ownership in Monkey Visual Cortex. *The Journal of Neuroscience*, 20(17):6594–6611.