Processing of Visual Statistics of Naturalistic Videos in Macaque Visual Areas V1 and V4

Gaku Hatanaka
Osaka University Graduate School of Frontier Biosciences: Osaka Daigaku Daigakuin Seimei Kino Kenkyuka

Mikio Inagakai
Osaka University Graduate School of Frontier Biosciences: Osaka Daigaku Daigakuin Seimei Kino Kenkyuka

Ryosuke F Takeuchi
Osaka University Graduate School of Frontier Biosciences: Osaka Daigaku Daigakuin Seimei Kino Kenkyuka

Shinji Nishimoto
Osaka University Graduate School of Frontier Biosciences: Osaka Daigaku Daigakuin Seimei Kino Kenkyuka

Koji Ikezoe
Yamanashi Daigaku - Kofu Campus: Yamanashi Daigaku

Ichiro Fujita (fujita@fbs.osaka-u.ac.jp)
Osaka University Graduate School of Frontier Biosciences: Osaka Daigaku Daigakuin Seimei Kino Kenkyuka https://orcid.org/0000-0003-3293-8610

Research Article

Keywords: two-photon microscopy, calcium imaging, Portilla-Simoncelli statistics, encoding model analysis, primary visual cortex, functional architecture

Posted Date: June 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-579596/v1

Version of Record: A version of this preprint was published at Brain Structure and Function on March 14th, 2022. See the published version at https://doi.org/10.1007/s00429-022-02468-z.

20210531

Processing of visual statistics of naturalistic videos in macaque visual areas V1 and V4

[a]Gaku Hatanaka, [a,b]Mikio Inagaki, [a]Ryosuke F Takeuchi, [a,b]Shinji Nishimoto, [a,b,c,*]Koji Ikezoe, and [a,b,*]Ichiro Fujita

[a]Graduate School of Frontier Biosciences, Osaka University, Suita, Osaka 565-0871, Japan
[b]Center for Information and Neural Networks, Osaka University and National Institute of Information and Communications Technology, Suita, Osaka 565-0871, Japan
[c]Faculty of Medicine, University of Yamanashi, Chuo, Yamanashi 409-3898, Japan

[*]**Corresponding authors**: Ichiro Fujita (fujita@fbs.osaka-u.ac.jp), Koji Ikezoe (kikezoe@yamanashi.ac.jp)

ORCID: Gaku Hatanaka (0000-0001-6724-8228), Mikio Inagaki (0000-0002-5294-501X), Ryosuke Takeuchi (0000-0003-4200-5259), Shinji Nishimoto (0000-0001-8015-340X), Koji Ikezoe (0000-0003-2736-5037), Ichiro Fujita (0000-0003-3293-8610)

Keywords: two-photon microscopy, calcium imaging, Portilla-Simoncelli statistics, encoding model analysis, primary visual cortex, functional architecture
Number of pages: 27
Number of figures and tables: 7 figures and 0 tables

Declarations

Acknowledgment

We thank Mahya Amano, Takanori Fukazawa, Kohei Iino, and Yusuke Saito for technical assistance, and Masashi Omokawa for animal care. This work was supported by the Japan Society for the Promotion of Science/Japanese Ministry of Education, Culture, Sports, Science, and Technology [KAKENHI grant numbers: JP15H01437, JP16H01673, JP17H01381, JP18H05007, JP21H02596 to I.F.; 15K18357 to K.I.; JP18H05522 to S.N.]; ERATO (JPMJER1801) to S.N; the National Institute of Information and Communications Technology; the Ministry of Internal Affairs and Communications. R.F.T. was supported by the Japan Society for the Promotion of Science Research Fellowship for Young Researchers.

Conflicts of interest/Competing interests

The authors declare no competing interests.

Ethics approval (include appropriate approvals or waivers)

All procedures were approved by the Animal Experiment Committee of Osaka University (permit numbers: FBS-07-017, FBS-12-017), and conformed to *the Guide for the Care and Use of Laboratory Animals* issued by the National Institutes of Health, USA.

Consent to participate (include appropriate statements): Not applicable.

Consent for publication (include appropriate statements): Consent for publication was obtained from all authors.

Availability of data and materials (data transparency)

The datasets for figures are available upon reasonable request to the corresponding authors. Visual stimuli (https://crcns.org/data-sets/vc/vim-2/abpit-vim-2) and part of the fluorescence data (https://ai-data.nict.go.jp/dataset/detail/?id=35) used in this study can also be found online.

Authors' contributions

G.H., M.I., R.F.T., and I.F. designed the research; G.H., R.T., and K.I. performed the research; G.H., M.I., and K.I. analyzed the data; and G.H., M.I., R.F.T., S.N., K.I., and I.F. wrote the paper.

Abstract

Natural scenes are characterized by diverse image statistics, including various parameters of the luminance histogram, outputs of Gabor-like filters, and pairwise correlations between the filter outputs of different positions, orientations, and scales (Potilla-Simoncelli statistics). Some of these statistics capture the response properties of visual neurons. However, it remains unclear to what extent such statistics can explain neural responses to natural scenes and how neurons that are tuned to these statistics are distributed across the cortex. By using two-photon calcium imaging and an encoding-model approach, we addressed these issues in macaque visual areas V1 and V4. For each imaged neuron, we constructed an encoding model to mimic its responses to naturalistic videos. By extracting Potilla-Simoncelli statistics through outputs of both filters and filter correlations, and by computing an optimally weighted sum of these outputs, the model successfully reproduced responses in a subpopulation of neurons. We evaluated the selectivities of these neurons by quantifying the contributions of each statistic to visual responses. Neurons whose responses were mainly determined by Gabor-like filter outputs (low-level statistics) were abundant at most imaging sites in V1. In V4, the relative contribution of higher-order statistics, such as cross-scale correlation, was increased, and the neuronal selectivities varied markedly across sites; many sites included numerous neurons sensitive to luminance histogram parameters and/or correlation statistics, whereas some sites were dominated by neurons responding to low-level statistics. The results indicate that natural scene analysis progresses from V1 to V4, and neurons sharing preferred image statistics are locally clustered in V4.

Introduction

The visual system performs complex analyses of input images derived from natural scenes. This complexity is highlighted by the fact that a wide variety of image statistics is necessary to synthesize artificial images that are perceptually indistinguishable from the original natural images (Freeman & Simoncelli, 2011). A subset of the statistics can be extracted by Gabor-like filters, which analyze the spatial frequency and orientation of local regions of the image (spectral statistics). Another subset of statistics is derived from computations of the correlations across the outputs of pairs of Gabor-like filters with different positions, orientations, and scales (correlation statistics). Also important in determining the appearance of an image are the summary statistics of the luminance histogram (i.e., luminance distribution), such as the mean, variance, skewness, and kurtosis (marginal statistics; Motoyoshi et al., 2007). The ensemble of these image statistics (hereafter referred to as Portilla-Simoncelli statistics, or PS statistics) was first proposed for texture analysis/synthesis (Portilla & Simoncelli, 2000), and its extension works well for explaining the perception of complex natural scenes by human observers (Freeman & Simoncelli, 2011).

Simple cells in the primary visual cortex (V1) have a linear receptive field with a Gabor-like structure and are sensitive to the phase of grating stimuli (Hubel and Wiesel 1962; Jones and Palmer 1987a, 1987b). Complex cells are tolerant to changes in the stimulus phase or stimulus position within their receptive field, and signal the local orientation, spatial frequency, motion, and binocular disparity independent of the sign of the stimulus contrast (Adelson & Bergen, 1985; Watson & Ahumada, 1985; Ohzawa et al., 1990). Spectral statistics are the major factors that explain the responses of both types of V1 neurons to texture images (Freeman et al., 2013). In contrast, correlation statistics are critical to eliciting responses to natural texture images in V2 (Freeman et al., 2013). In V4, which is a mid-tier stage along the ventral visual pathway, neuronal tuning to the correlation statistics and the skewness of the luminance histogram become more explicit than in V2 (Okazawa et al. 2017). Hierarchical processing along the ventral visual pathway appears to gradually create the representation of higher-order features of texture images.

A challenge in analyzing neuronal tuning to PS statistics is the high dimensionality of the visual stimuli. Unlike spectral statistics, which simply represent the spatial frequency and orientation, correlation statistics consist of the combination of filter outputs across

various positions, orientations, and scales, resulting in a large number of stimulus parameters. It is extremely difficult, if not impossible, to test all possible combinations of stimulus parameters in physiology experiments conducted in animals. Okazawa et al. (2015, 2017) mitigated this problem when analyzing the selectivity of V2 and V4 neurons for static texture stimuli by focusing on a subset of material surfaces and applying an adaptive sampling procedure (Yamane et al. 2008; Carlson et al., 2011). They successfully predicted the responses to texture images by means of a limited number of PS statistics. However, the contribution of marginal statistics to neuronal responses has been largely unexplored, because previous studies used stimuli with equalized luminance across stimulus images (Freeman et al., 2013; Okazawa et al. 2015, 2017). It remains unclear how well the responses to natural scenes can be characterized by PS statistics. Specifically, the luminance distribution is diverse across natural scene images, and therefore there are significant changes in the marginal statistics of dynamic scenes such as those in a video. As a step to extend our understanding of neuronal processing of the image statistics in more general natural scenes, we examined the effects of the PS statistics in naturalistic videos on the neuronal responses in V1 and V4.

In primates, neurons in V1 and V4 are spatially clustered based on their specific functional properties. Across the cortical surface of V1, neurons are arranged according to their selectivity to orientation (Hubel and Wiesel, 1968; 1977; Blasdel and Salama, 1986) and spatial frequency (Nauhaus et al., 2012). In V4, neurons with preferences for color, binocular disparity, orientation, curvature, and non-Cartesian gratings tend to be clustered locally (Gallant et al., 1996; Watanabe et al., 2002; Tanabe et al., 2005; Kotake et al., 2009; Tanigawa et al., 2010; Hu et al., 2020; Tang et al., 2020; Srinath et al., 2021). These findings suggest that V4 is functionally heterogeneous across the cortical surface in terms of visual stimulus selectivities (Roe et al., 2012). The functional architecture underlying various PS statistics parameters beyond orientation and spatial frequency has not been systematically explored in any visual cortical area. Here, we studied the spatial distribution of neurons sensitive to different PS statistics in V1 and V4. For this purpose, we used *in vivo* two-photon calcium imaging to record activities from a large number of neurons, the locations of which can be determined with high spatial resolution. An adaptive sampling procedure cannot be combined with simultaneous recordings from many neurons because it customizes the stimulus set through a genetic algorithm to suit the properties of a single target neuron. Instead, we applied an encoding-model approach (Gallant et al., 2011; Nishimoto et al., 2011) to

characterize neuronal tuning to PS statistics. This approach enables analysis on visual properties of hundreds of neurons simultaneously recorded by two-photon calcium imaging (Ikezoe et al., 2018).

We showed that an encoding-model analysis of two-photon imaging data can capture responses of a subpopulation of neurons to natural scenes. Consistent with previous findings on orientation and spatial frequency preferences, we found that neurons whose responses were determined largely by low-level statistics (spectral statistics) were dominant at most imaging sites in V1. In V4, the relative contribution of marginal and correlation statistics became greater. Neurons sensitive to luminance distribution and/or higher-order statistics such as linear scale correlation dominated many V4 sites, while some sites contained many neurons responding to low-level image statistics, suggesting that neurons processing different categories of image statistics are locally clustered in V4.

Materials and Methods

We performed two-photon calcium imaging in areas V1 and V4 of monkeys. All experimental and animal care procedures were approved by the animal experiment committee of Osaka University (permit numbers: FBS-12-017, FBS-18-005), and conformed to the *Guide for the Care and Use of Laboratory Animals* issued by the National Institutes of Health, USA (1996). The recording data for V1 largely overlapped with those presented in a previous study (Ikezoe et al., 2018), and here we subjected those data to novel analyses.

Animal preparation

We used two adult female monkeys (*Macaca fascicularis*, 2.3 and 3.2 kg). An initial aseptic surgery was performed for later repeat recordings (Ikezoe *et al*., 2013, 2018). Atropine sulfate (Mitsubishi Tanabe Pharma, Osaka; 0.025 mg/kg, intramuscular) was administered prior to each day's recording sessions. Then, anesthesia was induced with ketamine hydrochloride (Daiichi-Sankyo, Tokyo; 11.5 mg/kg, intramuscular), and maintained with isoflurane (1%–3%) in a mixture of nitrous oxide and oxygen (7:3). We performed a small craniotomy and durotomy (2–3 mm in diameter) over the region of V1 or V4 to be imaged. The exposed dura and cortex were covered with 2% agar and a 0.13- to 0.17-mm-thick coverglass (see Figure 1 of Ikezoe et al., 2013). After this surgical procedure, we switched from isoflurane to fentanyl citrate (Daiichi-Sankyo; 10 μg/kg/h, intravenous) for analgesia (Popilskis and Kohn, 1997). During imaging, vecuronium

bromide (Merck Sharp and Dohme, Tokyo; 0.08 mg/kg/h, intravenous) was given to the monkeys to prevent eye movements, and they were artificially ventilated with a mixture of nitrogen and oxygen. We maintained the body temperature at 37–38°C, and end-tidal CO_2 at 4.0%–5.5%. An electrocardiogram, blood pressure, and arterial oxygen-saturation levels were continuously monitored and maintained within appropriate ranges. Phenylephrine hydrochloride and tropicamide were applied to the eyes to relax accommodation and dilate the pupils. Corneas were covered with contact lenses of appropriate curvature, power, and pupil diameter (3 mm) to prevent drying and to allow the images on the stimulus display to be focused on the retina (Tamura et al., 2004). At the end of each recording session, the monkeys were administered neostigmine methylsulfate (Shionogi, Osaka; 0.1 mg/kg, intramuscular) to aid in recovering spontaneous respiration, as well as ketoprofen (Nissin Pharmaceutical, Yamagata, 0.8 mg/kg, intramuscular) for post-surgery analgesia. They were then returned to their cages and given ketoprofen and antibiotics (piperacillin sodium, Taisho Toyama Pharmaceutical, Tokyo) for 1 week after the experiment.

In vivo calcium imaging

We performed *in vivo* Ca^{2+} imaging of neuronal responses to naturalistic videos (see below) in V1 (17 sites) and V4 (16 sites). We pressure-injected a membrane-permeable fluorescent calcium indicator, Cal-520 AM (Cal-520, AAT Bioquest, Sunnyvale, CA; Tada et al., 2014), at a depth of 150–310 μm from the cortical surface through a micropipette. This depth corresponds to layer 2 and the uppermost tier of layer 3 (see Fig. 7 of Xu et al., 2003). Cal-520 was dissolved at a concentration of 1.0 mM in a solution containing 0.2% Pluronic F-127 (Thermo Fisher Scientific, Waltham, MA), 2.5% dimethyl sulfoxide (Sigma-Aldrich, St. Louis, MO), 10 mM HEPES, 2.5 mM KCl, and 150 mM NaCl (pH 7.4). We injected sulforhodamine 101 (SR 101; 100 μM, Thermo Fisher Scientific) together with Cal-520 to identify astrocytes (Nimmerjahn et al., 2004). A two-photon microscope (MOM, Sutter, Novato, CA) equipped with a mode-locked Ti:sapphire laser (MaiTai, Spectra Physics, Santa Clara, CA) was used for imaging fluorescence responses. The microscope was equipped with a 16× objective lens (CFI75 LWD 16× W, NA 0.8, Nikon, Tokyo), and controlled by a resonant scanner for the X direction, a galvano-scanner for the Y direction, and MOM Computer System and Software 2.0 (Sutter). The image sampling rate was 31 Hz. In V1, the imaged areas were 315 μm × 315 μm (eight sites), 355 μm × 355 μm (eight sites), or 630 μm × 630 μm (one

site). In V4, the imaged areas were 315 μm × 315 μm (seven sites), 355 μm × 355 μm (seven sites), or 630 μm × 630 μm (two sites).

Visual stimulation

We presented naturalistic videos (10° × 10°) on a gray background to the eye contralateral to the imaged hemisphere using a liquid crystal display (MDT231WG, Mitsubishi Electric, Tokyo) or an organic electroluminescence display (PVM-1741, Sony, Tokyo). The center portion of the videos covered the minimum response fields of the neurons that had been determined earlier by recording either multiple-unit electrical activities or fluorescence responses to presentations of 2° × 2° square patches of naturalistic videos at varying positions. The videos were composed of short clips (~10 s) showing different scenes including landscapes, animals, and humans, as well as letters and sentences (Nishimoto and Gallant, 2011; Nishimoto et al., 2011; available at https://crcns.org/data-sets/vc/vim-2/about-vim-2). The frame rate was 24 Hz. To train the encoding model, we prepared a 30-min video by concatenating the short clips to eliminate gaps, and then divided it into three 10-min videos (training sets, 10 min × three videos = 30 min). To test the quality of the model, we prepared a 3-min video that was not used for model training, and divided it into three 1-min videos. The videos were copied 10 times (1 min × three videos × 10 repetitions) and then pseudo-randomly reorganized into three 10-min videos (test sets, 10 min × three videos = 30 min). The 1-min videos appeared three or four times in each 10-min video. The training and test sets were interleaved with intervals of 1 to 10 min.

Calcium imaging processing

In the first image-processing step, we corrected constant image distortion caused by non-linear resonant scanning by mapping the pixels to correct positions using a template of the resonant scanner position against time. We then used two-dimensional cross-correlation to adjust image displacement caused by motion of the brain (Guizar-Sicairos et al., 2008). We quantified the fluorescence from individual neurons by selecting the cell-body pixels for each neuron from the average across all acquired images and then averaging the fluorescence across the cell-body pixels. We obtained the time courses of fluorescence responses by repeating this procedure for all sampled images. The fluorescence data were high-pass filtered with a cut-off frequency of 0.02 Hz to remove slow fluctuations of signal strength that are caused by movement of the cortex in depth direction or by gradual photobleaching over recording sessions. We normalized the

changes (dF) in fluorescence signals by the DC component (F_0) of the fluorescence ($\frac{dF}{F0} = \frac{F-F0}{F0}$). The DC component was estimated by applying the median filter to raw signals (filter width = 50 s).

We evaluated the consistency of responses across 10 test video presentations by calculating e*xplainable variance* (Sahani and Linden, 2003) as follows:

$$\bar{y}(t) = \frac{1}{N}\sum_{n=1}^{N} y_n(t)$$

$$Explainable\ variance = 1 - \frac{Var(y(t) - \bar{y}(t))}{Var(y(t))}$$

where $y_n(t)$ is the fluorescence signal (dF/F_0) for a time bin in the *n*-th trial, *N* is the number of trials, and $Var(\cdot)$ denotes the variance. Explainable variance indicates the proportion of variance of fluorescence signals $y(t)$ that can be explained by stimulus-related components $\bar{y}(t)$. An *explainable variance* of 1 indicates the same fluorescence responses across repeated presentations of an identical video. An explainable variance of 0 indicates inconsistent fluorescence responses across trials.

Extraction of Portilla-Simoncelli statistics from naturalistic videos

We decomposed an entire area of visual images of naturalistic videos into a set of PS statistics, each reflecting a different feature of a visual scene (Fig. 1a; Portilla and Simoncelli, 2000). We classified PS statistics into three groups: spectral, correlation, and marginal statistics. The spectral group of PS statistics captures the magnitudes of different orientations and spatial frequencies within an image. First, the videos were converted into the Commission International de l'Éclairage (CIE) L*A*B* color space, and only their luminance (L*) information was input into Gabor-like linear filters. Then energy filter outputs were produced by computing the amplitudes of the outputs of phase-shifted linear filter pairs. Thus, spectral statistics included only energy filter outputs averaged across an image. Correlation statistics consist of correlations of outputs of Gabor-like linear filters between different positions (linear position), orientations (linear orientation), and scales (linear scale), as well as correlations of outputs of energy filters between different positions (energy position), orientations (energy orientation), and scales (energy scale). Note that scale correlations were computed for combinations of different orientations as well as same orientations. Correlation statistics capture aspects

of the image structure such as angled or curved contours, sharp edges, and periodic textural patterns involving a combination of different orientations, spatial frequencies, or positions (Portilla and Simoncelli, 2000; Freeman et al., 2013). Marginal statistics refer to the summary statistics of the pixel luminance distribution of an image, such as the mean, variance, skewness, kurtosis, minimum value, and maximum value. Skewness and kurtosis were also computed for low-pass images at different cutoff frequencies (approximately ~18, ~9, ~4.5, and ~2.3 cycles/°; in order from higher to lower: skewL1, skewL2, skewL3, and skewL4; kurtL1, kurtL2, kurtL3, and kurtL4). For high-pass image (18 cycles/° ~), the variance was computed (VarHP).

We computed spectral and correlation statistics using six orientations (0°, 30°, 60°, 90°, 120°, and 150°), three spatial frequency bands (center frequencies, 12.8, 6.4, and 3.2 cycles/°; band width, approximately one octave), and 7 × 7 position shifts. Spectral statistics of high-pass (18 cycles/° ~) and low-pass (~2.3 cycles/°) images were also included without decomposing into orientation bands. A total of 936 PS statistics (spectral, 20; marginal, 15; linear position, 100; linear orientation, 60; linear scale, 174; energy position, 450; energy orientation, 45; energy scale, 72) were computed for a single video frame. We obtained the time courses of the PS statistics by concatenating them across video frames, and then compressed them to 6 Hz by averaging every four consecutive frames. Individual statistics were normalized by the mean and standard deviation, i.e., *z*-scored. To compute PS statistics, we used the code provided at the website of the Simoncelli laboratory (http://www.cns.nyu.edu/~lcv/texture/).

---------- Figure 1 comes near here ----------

Encoding-model analysis

We characterized the response selectivity of each neuron to a variety of PS statistics by fitting an encoding model to the fluorescence signals evoked by the naturalistic videos (Fig. 1). An array of linear temporal filters received outputs of the corresponding PS statistics filters (Fig. 1a). These temporal filters, spanning delays from 0 to 1 s at an interval of 1/6 s, were used to reproduce the delayed contributions of the PS statistics time courses to the fluorescence signals (Smetters et al., 1999; Ikezoe et al., 2018). The signals were then linearly summed across the filters to produce the model outputs.

We modeled fluorescence signals for individual neurons by optimizing temporal filters (Fig. 1a) at the second stage using L2-regularized linear regression (ridge regression; Huth et al., 2012). Fluorescence signals were normalized by the mean and standard deviation in each neuron to compensate for differences in signal strength across neurons. To optimize regularization parameters, we divided the entire 30-min training data into 50 chunks (each chunk = 36 s), and randomly picked 90% of the chunks for regression. For each recording site, we chose the regularization parameter that achieved the highest response-prediction accuracy for the remaining 10% of the data (the optimal regularization parameter). The final model was then obtained by regressions that included all training data and the optimal regularization parameter. The performance of the constructed model was tested using another video (test video; 3 min) that had not been used for model construction (Fig. 1b). Neuron-wise modeling accuracy was quantified using Pearson's correlation coefficient (r) between the measured and predicted fluorescence signals for the test set.

To quantify the relative contribution of marginal and correlation statistics over spectral statistics, we computed the non-spectral statistics ratio by dividing the mean absolute value of the second-stage filter weights of correlation statistics and marginal statistics by the mean absolute value of the weights of all PS statistics.

Results

Reliability of fluorescence responses of individual neurons to naturalistic videos

We recorded fluorescence responses to naturalistic videos from 2146 neurons in V1 (17 sites) and 2492 neurons in V4 (16 sites) of two monkeys. Fluorescence strength dynamically changed during the presentation of the videos. Figure 2a shows fluorescence responses of example neurons in V1 (#1 to #4) and V4 (#5 to #8). Gray traces represent responses of individual trials and black traces represent the trial average. Each neuron consistently responded to the videos at roughly the same time point across trials. To determine how well the fluorescence signals reflected neuronal stimulus selectivities rather than response fluctuations caused by internal and recording noise, we quantified the consistency of the fluorescence signals across trials by computing the explainable variance (Sahani and Linden, 2003; see Materials and Methods). It ranged from 0.029 to 0.77 in V1 and from 0.013 to 0.71 in V4 (Fig. 2b; medians were 0.16 for V1 and 0.12 for V4), meaning that 16% and 12% of the variance in the responses in V1 and V4,

respectively, could be explained by the variety of video images, and thus reflected neuronal visual selectivities. The explainable variance was larger in V1 than V4 (Wilcoxon rank-sum test, $p < 10^{-5}$), indicating that the signals were more robust against noise in V1 than in V4.

---------- Figure 2 comes near here ----------

Prediction performances of PS models

For each neuron, we constructed an encoding model based on PS statistics to characterize the neuronal responses to naturalistic videos. We sought to simulate responses to a training video of 30 min by defining the optimal set of weights for PS statistics (Fig. 1a). The performance of the constructed model was tested using another video (test video; 3 min) that had been kept unused for model construction (Fig. 1b). The model response to the test video varied along the time axis as the contents of the video changed (Fig. 3a; V1, green; V4, orange). The time course of the model output depended on the visual selectivity of the corresponding neurons, and thus differed across the models. For the successful example neurons shown in Figure 3a (#2674 in V1, #567 in V4), the outputs of the models faithfully followed the fluorescence responses of the corresponding neurons (Fig. 3a, black). Note that neuronal responses were averaged across 10 trial repetitions to minimize the effects of noise. The prediction performances of the models as evaluated by Pearson's correlation coefficients between the model outputs and the neuronal responses were 0.61 for the V1 neuron (#2674) and 0.62 for the V4 neuron (#567). In less successful examples (#2673 in V1, #1943 in V4), the model outputs moderately followed the neuronal responses and the correlation coefficients were 0.35 (#2673) and 0.30 (#1943). The mean correlation coefficients across neurons were 0.19 ± 0.13 (n = 2146; ± s.d.) in V1 and 0.11 ± 0.11 (n = 2492) in V4. The distributions of the correlation coefficients were broad in both areas (Fig. 3b), indicating that some models predicted the responses relatively accurately but other models did so only moderately or poorly. The prediction performance was positively correlated with the explainable variance (Spearman's correlation coefficient $rs = 0.39$ in V1, $rs = 0.18$ in V4; $p < 0.0001$), indicating that as expected, the models worked better for neurons with more robust responses across trials. Despite the variety across the models, on a population-wide level the correlation coefficient was significantly higher than zero both in V1 (Wilcoxon signed-rank test, $p < 10^{-5}$) and V4 ($p < 10^{-5}$). Our encoding-model approach captured the visual selectivity of neuronal subpopulations by means of PS statistics. The

correlation coefficient was higher in V1 than V4 (Wilcoxon rank-sum test, $p < 10^{-5}$). For the following analyses, we chose the neurons with correlation coefficients exceeding 0.2 (dark columns in Fig. 3b; 972 neurons in V1, 477 neurons in V4; Ikezoe et al., 2018). Changing this selection criterion from 0.2 to 0.1 or 0.3 did not affect the overall conclusions of this study. The sizes of receptive fields of these neurons, estimated as the standard deviation of two-dimensional Gaussian fitting to the response map in *in-silico* simulations (Nishimoto et al., 2011; Ikezoe et al., 2018), were 1.13° ± 0.80° (mean ± s.d.) in V1 and 2.28° ± 1.91° in V4. The receptive fields were twice as large in V4 as in V1 ($p < 10^{-5}$; Wilcoxon rank-sum test).

---------- Figure 3 comes near here ----------

Comparison of tunings for image statistics in V1 and V4

We examined how neuronal tunings to PS statistics contributed to responses to naturalistic videos. In this analysis, we grouped the PS statistics into eight categories: spectral, marginal, and correlations between linear positions, linear orientation, linear scales, energy positions, energy orientations, and energy scales (see Materials and Methods). Each category contained a number of statistics, and each statistic had its own fitting weight. For each category, we chose the largest fitting weight (i.e., the highest peak of the temporal filters) of constituent statistics to represent their importance in the neuronal responses (Fig. 4a).

In many neurons in V1 (42%, 411/972 neurons), spectral statistics had the largest fitting weight (Fig. 4a, upper panel), as we would expect given the well-known tunings of V1 neurons to spatial frequency and orientation. Marginal statistics were critical for responses of another large population of V1 neurons (41%, 399/972). In a small subset of V1 neurons, linear scale correlation had the largest weight (16%, 153/972). The responses of 1% of V1 neurons (nine neurons) were best explained by linear orientation correlation. All other groups of statistics, such as linear position correlation and the three energy correlations, contributed little to the responses of V1 neurons.

In contrast, more than half of V4 neurons were influenced by marginal statistics (54%, 257/477). The proportion of neurons in which spectral statistics had the largest weight (22%, 103/477) was relatively small compared to V1 (Fig. 4a, lower panel). In addition, unlike V1 neurons, many of the V4 neurons that were tuned to spectral statistics were

also influenced by higher-order statistics, including energy correlations; the weights of these correlations were larger than the half-maximum weight (see bright colors in the rows for energy position, energy orientation, and energy scale in Fig. 4a, lower panel). In a smaller population of V4 neurons, linear scale correlation had the largest weight (22%, 106/477); the percentage of these neurons was slightly higher than that in V1 (16%). A minority of V4 neurons were preferentially tuned to other linear or energy correlations (2.3%, 11/477).

---------- Figure 4 comes near here ----------

The overall profiles of the fitting weight across the eight categories of PS statistics were similar between V1 and V4 (Fig. 4b). The spectral and marginal statistics had the largest weights, followed by linear scale correlation. More detailed comparisons between V1 and V4 revealed that the weight of spectral statistics was larger, on average, in V1 than in V4 (Wilcoxon rank-sum test, Bonferroni correction; $p < 10^{-5}$). In contrast, the weight of marginal statistics was larger in V4 than in V1 ($p = 0.013$, Fig. 4b). The weights of correlation statistics were generally larger in V4 than in V1. Specifically, the weights of the linear position correlation ($p = 7.0 \times 10^{-3}$), energy position correlation ($p < 10^{-5}$), energy orientation correlation ($p < 10^{-5}$), and energy scale correlation ($p < 10^{-5}$) were significantly larger in V4 than in V1. The weight of the linear scale correlation was the largest among the correlation statistics in V1 and V4 and was comparable between the two areas ($p > 0.05$).

We computed non-spectral statistics ratios to quantify how marginal and correlation statistics compared to spectral statistics in terms of their contribution to neuronal responses. The ratios were higher in V4 than in V1 (V4, median = 0.34, interquartile range = 0.11, n = 477; V1, 0.24 and 0.095, n = 972; Wilcoxon rank-sum test, $p < 10^{-5}$). These results demonstrate that tunings to marginal and correlation statistics were more explicit in V4 than in V1.

Because marginal statistics contain a variety of parameters (i.e., mean, variance, skewness, kurtosis, and range), we next analyzed the weight strength of each parameter (Fig. 5). The weights of the mean of the luminance histogram (Mean) were comparable between V1 and V4 (Wilcoxon rank-sum test, Bonferroni correction; $p > 0.05$ corrected for multiple comparisons between 15 marginal statistics), and larger than those of the

other parameters in each area. The weights of kurtosis values computed from the original (Kurt, $p < 10^{-5}$) and sub-band images (KurtL1, $p < 10^{-5}$; KurtL2, $p < 10^{-5}$; KurtL3, $p < 10^{-5}$; KurtL4, $p < 10^{-5}$) were larger in V1 than in V4, whereas the weights of the variance (Var, $p < 10^{-5}$) and some skewness values computed from coarser sub-bands (SkewL3, $p = 1.1 \times 10^{-5}$; SkewL4, $p < 10^{-5}$) were larger in V4 than in V1. No differences were found between V1 and V4 regarding skewness from the original (Skew, $p > 0.05$), maximum (Max, $p > 0.05$), skewness computed from finer sub-bands (SkewL1, $p > 0.05$; SkewL2, $p > 0.05$), and variance of high-pass filtered images (VarHP, $p > 0.05$).

---------- Figure 5 comes near here ----------

Spatial distribution of neurons with different tuning preferences

We next examined the spatial distribution of the neurons across the cortical surfaces of V1 and V4 according to their tunings to PS statistics. By labeling each neuron according to the category with the largest fitting weight, we created a map showing which PS statistic groups contributed most in individual neurons. In an example V1 site, nearly half of the neurons (43%, 15/35) were tuned to spectral statistics, while another 40% (14/35) of neurons were tuned to marginal statistics (site #1; Fig. 6a, b). The two groups of neurons were intermingled with each other. In another V1 site (site #2), the vast majority of neurons (79%, 41/52) were tuned to spectral statistics and only a small population (17%, 9/52) were tuned to marginal statistics (Fig. 6d, e). Although 16% of V1 neurons were tuned to linear scale correlation (Fig. 4a, upper panel), we found no sites in which these neurons were predominant.

The distribution of neurons with different tuning preferences varied across the recording sites more drastically in V4 than in V1. In site #3 (Fig. 6g, h), most neurons (90%, 46/51) were tuned to marginal statistics and the rest were tuned to linear scale correlation. Few neurons at this site were sensitive to spectral statistics. In contrast, in another V4 recording site, most neurons (93%, 53/57) were preferentially tuned to spectral statistics (site #4, Fig. 6j, k). Note that these neurons were also sensitive to correlation statistics, and the weights of several subgroups of correlation statistics reached half the value of the largest weight of spectral statistics (Fig. 6k). The non-spectral statistics ratios in V4 sites were relatively high (median: 0.37, Fig. 6l; 0.35, Fig. 6i), indicating that the responses of the neurons at these sites were influenced by marginal and correlation statistics in addition to spectral statistics. In contrast, the non-spectral statistics ratios were relatively

low in the V1 sites (median, 0.25; Fig. 6c, 0.25; Fig 6f, 0.19), confirming that simple features defined by spectral statistics contributed more to neuronal responses in V1 than in V4.

---------- Figure 6 comes near here ----------

The variability of the tuned PS statistics categories across imaging sites is evident in a summary figure illustrating the percentages of neurons tuned to spectral, marginal, and correlation statistics in all imaging sites (Fig. 7). In V1, neurons tuned to spectral statistics (green bars) and those tuned to marginal statistics (grey bars) were dominant, with varying relative proportions across sites. Neurons tuned to correlation statistics (bars in blue and other colors) were a minor population in all V1 sites. In V4, neurons tuned to correlation statistics constituted more than 50% of the imaged neurons in seven sites, and less than 30% in the remaining nine sites. In two sites in V4, neurons tuned to spectral statistics were predominant (the lowest two rows in Fig. 7b). These results suggest that neurons with selectivities for different PS statistics groups were not uniformly distributed, but rather were locally clustered in both V1 and V4. In particular, V4 consisted of sites dominated by neurons tuned to correlation statistics and sites with only a small proportion of such neurons.

---------- Figure 7 comes near here ----------

Clustering of neurons with similar selectivities to PS statistics was also supported by the relation between intercellular distance and response similarities. As the distance between the neuron pair became larger, pairwise correlation of fitting weights of PS statistics became smaller in V1 and V4 (V1, $rs = -0.22$; $p < 10^{-5}$; V4, $rs = -0.22$, $p < 10^{-5}$; test with 100000 permutations). Neighboring neurons tended to exhibit similar tuning to PS statistics, suggesting that neurons with similar tuning properties were locally clustered.

Discussion

By combining two-photon calcium imaging and encoding-model analysis, we characterized the response selectivities of neurons in V1 and V4 to PS statistics based on their responses to naturalistic videos. The responses of many neurons in both areas were explained most by the spectral statistics, marginal statistics, or linear scale correlation in the video images. Overall, the responses of V4 neurons depended on marginal statistics

and correlation statistics to a greater extent than the responses of V1 neurons. We further showed that neurons with different selectivities to PS statistics were distributed unevenly within each area, and neighboring neuron pairs exhibited more similar selectivities to PS statistics than distant neuron pairs. Of particular note is that different imaging sites in V4 showed drastically different proportions of neurons tuned to spectral, marginal, or correlation statistics, suggesting that V4 contained discrete subregions with different selectivities to PS statistics.

Encoding-model approach to studies of neuronal tuning

We exploited an encoding-model approach to analyze the effect of a large number of PS statistics and to minimize the overfitting problem by incorporating regularization. The prediction performance of the models varied greatly across neurons, but the distribution deviated from zero towards positive values (Fig. 3), indicating that the encoding model was able to capture the responses to the naturalistic videos in a significant subpopulation of V1 and V4 neurons. The prediction performance of the encoding models was higher for V1 neurons than for V4 neurons (Fig. 3b). The positive correlation between explainable variance and prediction performance suggests that the higher performance of the V1 models may be at least partly due to the more consistent responses of V1 neurons across trials (Fig. 2b).

The use of naturalistic videos as visual stimuli has both merits and demerits. On one hand, this approach allows us to characterize neuronal selectivities to marginal statistics that change dynamically during natural vision. We found that marginal statistics were crucial in determining the responses of a large population of neurons in V1 and V4 (Fig. 4a, b), and these responses were affected by many aspects of the luminance distribution, such as the mean, variances, skewness, and kurtosis (Fig. 5). This sensitivity to marginal statistics has been underestimated in previous studies using stimuli with equalized marginal statistics across images. We implemented PS-statistics filters as the initial filters in our models, because the ventral visual pathway areas process shape and texture that can be synthesized from PS-statistics-based computation (Portilla and Simoncelli, 2000; Freeman and Simoncelli, 2011). On the other hand, our PS-statistics–based models cannot capture responses to other visual cues in the videos, such as color and motion. Given the selectivies of V1 and V4 neurons to color (Komatsu, 1998; Gegenfurtner 2003; Roe et al., 2012), it will be interesting to see if the addition of color detection filters to the models improves prediction performance. An encoding model that used a motion energy

model (Ikezoe et al., 2018) achieved higher prediction performance in V1 (mean ± s.d., 0.36 ± 0.15, n = 2146) but comparable performance in V4 (0.16 ± 0.12, n = 2492) compared to performances with the PS-statistics model in V1 (0.19 ± 0.13) and V4 (0.11 ± 0.11). This finding suggests that motion plays an important role in determining the responses of the V1 neurons in our sample, i.e., the V1 neurons in layer 2 and the uppermost tier of layer 3.

The adaptive stimulus sampling procedure customizes the stimulus set to the properties of the targeted individual neuron, thus making it possible to achieve higher prediction accuracy (Okazawa et al., 2015). Our encoding-model analysis does not include a stimulus optimization procedure for individual neurons, but instead is applicable to simultaneously recorded responses from multiple neurons. Using two-photon calcium imaging, we were able to characterize the selectivities of several tens to hundreds of neurons in one experiment, resulting in a database consisting of up to a few thousand neurons in total, a substantially larger number than previously examined. Thus, there is a trade-off between model accuracy, achieved with single-neuron recording and the adaptive stimulus sampling procedure, and larger database size, which can be realized using two-photon calcium imaging of multiple neurons with the encoding-model approach. The information derived from our large database can complement the results of previous studies.

Tunings to PS statistics in V1 and V4

In V1 and V4, spectral statistics, marginal statistics, and linear scale correlation were major factors that explained the responses to the naturalistic videos (Fig. 4). Tuning to spectral statistics is consistent with the well-known selectivity of V1 and V4 neurons to the orientations and spatial frequencies of stimuli (e.g., for V1, Campbell et al. 1969; De Valois et al. 1982; Hubel and Wiesel 1962; Maffei and Fiorentini 1973; Movshon et al. 1978; for V4, Desimone and Schein 1987; Roe et al., 2012; Lu et al., 2018). We found that V4 neurons preferring spectral statistics tended also to be tuned to correlation statistics, including both linear and energy correlations between different scales, orientations, and positions (Fig. 4A). Correlation between different scales is useful for representing edges, correlation between different orientations is useful for representing curves and angles, and correlation between different positions is useful for representing repeated patterns such as visual texture. V4 neurons thus have properties suitable to respond to complex features such as angled or curved contours, sharp edges, and periodic

patterns of visual images. Previous studies have shown that a substantial population of V4 neurons respond to curved contours and gratings (Gallant et al., 1996; Pasupathy and Connor, 1999), corners and crosses (Hegdé and Van Essen, 2007), visual textures (Okazawa et al., 2015, 2017), and complex shapes (Kobatake and Tanaka, 1994; Carlson et al., 2011).

We showed that 41% of V1 neurons and 54% of V4 neurons were preferentially tuned to marginal statistics (Fig. 4), whereas previous studies found only a small number of such neurons (<10%; see Fig. 5C of Okazawa et al., 2017). The abundance of neurons tuned to marginal statistics was revealed by using naturalistic videos as stimuli. The contents of clips and frames in these videos drastically differed, and no normalization was applied to compensate for large differences in the luminance histograms across clips and frames. The marginal statistics computed from the luminance histograms therefore dynamically changed along the time axis. In contrast, previous studies equalized the mean and variance of luminance between their stimuli (Okazawa et al., 2015, 2017). Our results indicate that marginal statistics, including skewness as well as the mean and variance of luminance, are important factors in determining the responses to natural scenes in both V1 and V4 (Fig. 5). Indeed, some V1 neurons selectively respond to the luminance level of uniform surfaces that lack spectral statistics (Maguire & Baizer, 1982; Rossi et al., 1996; Kinoshita and Komatsu, 2001). Visual features unrelated to spectral and correlation statistics are likely to be represented in V1 and further processed in V4.

Linear scale correlation underlies the representation of visual features produced by interactions across different spatial frequencies. For example, edge-like and line-like features are produced by the same orientation, but with different phase congruence across spatial frequencies. Selective responses to such phase congruence were found in V1 neurons in studies using compound gratings containing multiple spatial frequencies (Mechler et al., 2002). The neuronal tuning to linear scale correlation that we observed in V1 and V4 neurons might explain the selectivity to the spatial phase congruence.

Comparisons of fitting weights revealed that spectral statistics contributed more to neuronal responses in V1 than in V4 (Fig. 4b). In contrast, the contribution of marginal and correlation statistics was greater in V4 than in V1 (Fig. 4c). These results are consistent with those of previous studies supporting the gradual development of neural

representations for higher-order features along the ventral visual pathway (Freeman et al., 2013; Okazawa et al., 2017).

Functional subregions in V4

In V1 and V4, the proportions of the preferred statistics depended on the imaging sites (Figs. 6, 7). In V1, the proportions of neurons tuned to spectral statistics and those tuned to marginal statistics varied gradually across sites, while the proportion of neurons tuned to correlation statistics remained consistently low (Fig. 7a). The differential distributions of neurons with different stimulus selectivities were more prominent in V4 (Fig. 7b); some sites contained a high proportion (50–70%) of neurons tuned to correlation statistics, whereas other sites contained a much lower proportion (10–30%) of such neurons, and a few sites consisted predominantly of neurons tuned to spectral statistics. This finding is further evidence for the existence of functional subregions within V4 as reported by studies using functional MRI and intrinsic signal optical imaging. Activation caused by color or orientation of stimulus gratings appears as separate spots in V4 (Conway et al., 2007; Tanigawa et al., 2010). The V4 sites with many neurons tuned to spectral statistics may be located in orientation-selective subregions. Further, the V4 sites containing neurons tuned to correlation statistics may preferentially respond to visual texture (Okazawa et al., 2017). Although clustering of neurons with texture sensitivity has not been examined to date, Gallant et al. (1996) previously suggested clustering of neurons that respond to similar non-Cartesian (curved) gratings. A recent study using intrinsic signal optical imaging demonstrated that curvatures of different degrees were processed in sub-millimeter-sized modules in V4 (Hu et al., 2020). Single-neuron recording studies also suggest that neurons with sensitivities to binocular disparity and disparity-defined edges are locally clustered in V4 (Watanabe et al., 2002; Tanabe et al., 2005; Fang et al., 2019). The combined application of macroscopic mapping using fMRI or intrinsic signal optical imaging with microscopic mapping using two-photon calcium imaging may shed light on the overall organization and finer details of these functional subregions of V4.

References

Adelson E, Bergen J (1985) Spatiotemporal energy models for the perception of motion. J Opt Soc Am A 2:284-299. https://doi.org/10.1364/JPSAA.2.000284

Blasdel GG, Salama G (1986) Voltage-sensitive dyes reveal a modular organization in monkey striate cortex. Nature 321:579-585. https://doi.org/10.1038/321579a0

Campbell FW, Cooper GF, Enroth-Cugell C (1969) The spatial selectivity of the visual cells of the cat. J Physiol 203:223-235. https://doi.org/10.1113/jphysiol.1969.sp008861.

Carlson ET, Rasquinha RJ, Zhang K, Connor CE (2011) A sparse object coding scheme in area V4. Curr Biol 21:288-293. https://doi.org/10.1016/j.cub.2011.01.013

Conway BR, Moeller S, Tsao DY (2007) Specialized color modules in macaque extrastriate cortex. Neuron 56:560-573. https://doi.org/10.1016/j.neuron.2007.10.008

Dai J, Wang Y (2012) Representation of surface luminance and contrast in primary visual cortex. Cerebral Cortex 22:776-787. https://doi.org/10.1093/cercor/bhr133

Desimone R, Schein SJ (1987) Visual properties of neurons in area V4 of the macaque: sensitivity to stimulus form. J Neurophysiol 57:835-868. https://doi.org/10.1152/jn.1987.57.3.835

De Valois RL, Albrecht DG, Thorell LG (1982) Spatial frequency selectivity of cells in macaque visual cortex. Vision Res 22:545-559. https://doi.org/10.1016/0042-6989(82)90113-4

Fang Y, Chen M, Xu H, Li P, Han C, Hu J, Zhu S, Ma H, Lu HD (2019) An orientation map for disparity-defined edges in area V4. Cerebral Cortex 29:666-679. https://doi.org/10.1093/cercor/bhx348

Freeman J, Simoncelli EP (2011) Metamers of the ventral stream. Nat Neurosci 14:1195-1201. https://doi.org/10.1038/nn.2889

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. https://doi.org/10.1038/nn.3402

Gallant JL, Connor CE, Rakshit S, Lewis JW, Van Essen DC (1996) Neural responses to polar, hyperbolic, and cartesian gratings in area V4 of the macaque monkey. J Neurophysiol 76:2718-2739. https://doi.org/10.1152/jn.1996.76.4.2718

Gallant JL, Nishimoto S, Naselaris T, Wu MCK (2011) System identification, encoding models and decoding models: a powerful new approach to fMRI research. In

Kriegeskorte N and Kreiman G (eds.), *Visual Population Codes* (pp.163-188), Cambridge: MIT press. https://doi.org/10.7551/mitpress/8404.003.0010

Gegenfurtner KR (2003) Cortical mechanisms of color vision. Nat Rev Neurosci 4:563-572. https://doi.org/10.1038/nrn1138

Guizar-Sicairos M, Thurman ST, Fienup JR (2008) Efficient subpixel image registration algorithms. Opt Lett 33:156-158. https://doi.org/10.1364/ol.33.000156

Hegdé J, Van Essen DC (2007) A comparative study of shape representation in macaque visual areas V2 and V4. Cerebral Cortex 17:1100-1116. https://doi.org/10.1093/cercor/bhl020

Hu J, Song XM, Wang Q, Roe AW (2020) Curvature domains in V4 of macaque monkey. eLife 9:e57261. https://doi.org/10.7554/eLife.57261:1-21.

Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol (Lond) 160:106-154. https://doi.org/10.1113/jphysiol.1962.sp006837

Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. J Physiol (Lond) 195: 215-243. https://doi.org/10.1113/jphysiol.1968.sp008455

Hubel DH, Wiesel TN (1977) Ferrier lecture: Functional architecture of macaque visual cortex. Proc R Soc B Biol Sci 198: 1-59. https://doi.org/10.1098/rspb.1977.0085

Huth AG, Nishimoto S, Vu AT, Gallant JL (2012) A continuous semantic space describes the representation of thousands of objects and action categories across the human brain. Neuron 76:1210-1224. https://doi.org/10.1016/j.neuron.2012.10.014

Ikezoe K, Amano M, Nishimoto S, Fujita I (2018) Mapping stimulus feature selectivity in macaque V1 by two-photon Ca^{2+} imaging: encoding-model analysis of fluorescence responses to natural movies. Neuroimage 180 (Pt A):312-323. https://doi.org/10.1016/j.neuroimage.2018.01.009

Ikezoe K, Mori Y, Kitamura K, Tamura H, Fujita I (2013) Relationship between the local structure of orientation map and the strength of orientation tuning of neurons in monkey V1: a 2-photon calcium imaging study. J Neurosci 33:16818-16827. https://doi.org/10.1523/JNEUROSCI.2209-13.2013

Jones JP, Palmer LA (1987a) The two-dimensional spatial structure of simple receptive fields in cat striate cortex. J Neurophysiol 58:1187-1211. https://doi.org/10.1152/jn.1987.58.6.1187

Jones JP, Palmer LA (1987b) An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J Neurophysiol 58:1233-1258. https://doi.org/10.1152/jn.1987.58.6.1233

Kinoshita M, Komatsu H (2001) Neural representation of the luminance and brightness of a uniform surface in the macaque primary visual cortex. J Neurophysiol 86:2559-2570. https://doi.org/10.1152/jn.2001.86.5.2559

Kobatake E, Tanaka K (1994) Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. J Neurophysiol 71:856-867. https://doi.org/10.1152/jn.1994.71.3.856

Komatsu H (1998) Mechanisms of central color vision. Curr Opin Neurobiol 8:503-508. https://doi.org/10.1016/s0959-4388(98)80038-x

Kotake Y, Morimoto H, Okazaki Y, Fujita I, Tamura H (2009) Organization of color-selective neurons in macaque visual area V4. J Neurophysiol 102:15-27. https://doi.org/10.1152/jn.90624.2008

Lu Y, Yin J, Chen Z, Gong H, Liu Y, Qian L, Li X, Liu R, Andolia IM, Wang W (2018) Revealing detail along the visual hierarchy: neural clustering preserves acuity from V1 to V4. Neuron 98:417-428. https://doi.org/10.1016/j.neuron.2018.03.009

Maffei L, Fiorentini A (1973) The visual cortex as a spatial frequency analyzer. Vision Res 13:1255-1267. https://doi.org/10.1016/0042-6989(73)90201-0

Maguire WM, Baizer JS (1982) Luminance coding of briefly presented stimuli in area 17 of the rhesus monkey. J Neurophysiol 47:128-137. https://doi.org/10.1152/jn.1982.47.1.128

Mechler F, Reich DS, Victor JD (2002) Detection and discrimination of relative spatial phase by V1 neurons. J Neurosci 22:6129-6157. https://doi.org/10.1523/JNEUROSCI.22-14-06129.2002

Motoyoshi I, Nishida S, Sharan L, Adelson EH (2007) Image statistics and the perception of surface qualities. Nature 447:206-209. https://doi.org/10.1038/nature05724

Movshon JA, Thompson ID, Tolhurst DB (1978) Spatial summation in the receptive fields of simple cells in the cat's striate cortex. J Physiol (Lond) 283:53-77. https://doi.org/10.1113/jphysiol.1978.sp012488

Naselaris T, Kay KN, Nishimoto S, Gallant JL (2011) Encoding and decoding in fMRI. Neuroimage 56:400-410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Nauhaus I, Nielsen KJ, Disney JJ, Callaway EM (2012) Orthogonal micro-organization of orientation and spatial frequency in primate primary visual cortex. Nat Neurosci 15:1683-1690. https://doi.org/10.1038/nn.3255

Nimmerjahn A, Kirchhoff F, Kerr JN, Helmchen F (2004) Sulforhodamine 101 as a specific marker of astroglia in the neocortex in vivo. Nat Methods 1:31-37. https://doi.org/10.1038/nmeth706

Nishimoto S, Gallant J (2011) A three-dimensional spatiotemporal receptive field model explains responses of area MT neurons to naturalistic movies. J Neurosci 31:14551-14564. https://doi.org/10.1623/JNEUROSCI.6801-10.2011

Nishimoto S, Vu AT, Naselaris T, Benjamin Y, Yu B, Gallant JL (2011) Reconstructing visual experiences from brain activity evoked by natural movies. Curr Biol 21:1641-1646. https://doi.org/10.1016/j.cub.2011.08.031

Ohzawa I, DeAngelis GC, Freeman RD (1990) Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. Science 249:1037-1041. https://doi.org/10.1126/science.2396096

Okazawa G, Tajima S, Komatsu H (2015) Image statistics underlying natural texture selectivity of neurons in macaque V4. Proc Natl Acad Sci USA 112:E351-E360. https://doi.org/10.1073/pnas.1415146112

Okazawa G, Tajima S, Komatsu H (2017) Gradual development of visual texture-selective properties between macaque areas V2 and V4. Cerebral Cortex 27:4867-4880. https://doi.org/10.1093/cercor/bhw282

Pasupathy A, Connor CE (1999) Responses to contour features in macaque area V4. J Neurophysiol 82:2490-2502. https://doi.org/10.1152/jn.1999.82.5.2490

Popilskis SJ, Kohn DF (1997) Anesthesia and analgesia in nonhuman primates. In: Kohn DF, Wixson SK, White WJ, Benson GJ (Eds), Anesthesia and analgesia in laboratory animals. Academic Press, San Diego, pp 233-255. https://doi.org/10.1016/B978-012417570-9/50014-3

Portilla J, Simoncelli EP (2000) A parametric texture model based on joint statistics of complex wavelet coefficients. Int J Comput Vis 40:49-70. https://doi.org/10.1023/A:1026553619983

Roe A, Chelazzi L, Connor CE, Conway BR, Fujita I, Gallant JL, Lu H, Vanduffel W (2012) Toward a unified theory of visual area V4. Neuron 74:12-29. https://doi.org/10.1016/j.neuron.2012.03.011

Rossi AF, Rittenhouse CD, Paradiso MA (1996) The representation of brightness in primary visual cortex. Science 273:1104-1107. https://doi.org/10.1126/science.273.5278.1104

Sahani M, Linden JF (2003) How linear are auditory cortical responses? In: Becker S, Thrun S, Obermayer K (Eds), Advances in neural information processing systems,

vol. 15. MIT Press, Cambridge, pp 109-116. http://www.gatsby.ucl.ac.uk/~maneesh/papers/nips02-linearity.pdf

Smetters D, Majewska A, Yuste R (1999) Detecting action potentials in neuronal populations with calcium imaging. Methods 18:215-221. https://doi.org/10.1006/meth.1999.0774

Srinath R, Emonds A, Wang Q, Lempel AA, Dunn-Weiss E, Connor CE, Nielsen KJ (2021) Early emergence of solid shape coding in natural and deep network vision. Curr Biol 31:51-65. https://doi.org/10.1016/j.cub.2020.09.076

Tada M, Takeuchi A, Hashizume M, Kitamura K, Kano M (2014) A highly sensitive fluorescent indicator dye for calcium imaging of neural activity *in vitro* and *in vivo*. Eur J Neurosci 39:1720-1728. https://doi.org/10.1111/ejn.12476

Tamura H, Kaneko H, Kawasaki K, Fujita I (2004) Presumed inhibitory neurons in the macaque inferior temporal cortex: visual response properties and functional interactions with adjacent neurons. J Neurophysiol 91:2782-2796. https://doi.org/10.1152/jn.01267.2003

Tanabe S, Doi T, Umeda K, Fujita I (2005) Disparity-tuning characteristics of neuronal responses to dynamic random-dot stereograms in macaque visual area V4. J Neurophysiol 94:2683-2699. https://doi.org/10.1152/jn.00319.2005

Tang R, Song Q, Li Y, Zhang R, Cai X, Lu HD (2020) Curvature-processing domains in primate V4. eLife 9:e57502:1-21. https://doi.org/10.7554/eLife.57502

Tanigawa H, Lu HD, Roe AW (2010) Functional organization for color and orientation in macaque V4. Nat Neurosci 13:1542-1548. https://doi.org/10.1038/nn.2676

Xu L, Tanigawa H, Fujita I (2003) Distribution of α-amino-3-hydroxy-5- methyl-4isoxazolepropionate-type glutamate receptor subunits (GluR2/3) along the ventral visual pathway in the monkey. J Comp Neurol 456:396-407. https://doi.org/10.1002/cne.10538

Watanabe M, Tanaka H, Uka T, Fujita I (2002) Disparity-selective neurons in area V4 of macaque monkeys. J Neurophysiol 87:1960-1973. https://doi.org/10.1152/jn.00780.2000

Watson A, Ahumada A (1985) Model of human visual-motion sensing. J Opt Soc Am 2:322-341. https://doi.org/10.1364/jossa.2.000322

Yamane Y, Carlson ET, Bowman KC, Wang Z, Connor CE (2008) A neural code for three-dimensional object shape in macaque inferotemporal cortex. Nat Neurosci 11:1352-1360. https://doi.org/10.1038/nn.2202

Figure Captions

Fig. 1 Encoding-model with first-stage filters computing Portilla-Simoncelli statistics

(a) The first stage of the encoding model consists of a set of spectral filters, correlations between outputs of different filters, and the marginal statistics of the luminance distributions of images (PS statistics). These filters produce time-varying outputs when a video is fed as an input. Each of these filters is followed by a temporal filter at the second stage. The model then linearly sums the second-stage filter outputs to produce the model response. For each neuron recorded, we trained the model by optimizing the temporal filters to mimic the neuron's fluorescence responses to a training video. **(b)** The model performance was evaluated by comparing the model response and the neuron response to a test video that was not used for the model construction. Encircled asterisks, convolution; encircled crosses, correlation; encircled plus sign, linear summation.

Fig. 2 Reliability and diversity of fluorescence responses of V1 and V4 neurons to naturalistic videos

(a) Fluorescence response time courses of example neurons in V1 (#1–#4) and V4 (#5–#8). Black lines indicate average responses across 10 trials, and gray lines indicate responses in individual trials. Numbers in the two-photon microscopy images correspond to the neuron numbers to the left of the fluorescence traces. In contrast with astroglia, neurons were sulforhodamine negative. dF/F denotes changes (dF) in fluorescence signals normalized by the DC component of the fluorescence (F_0). **(b)** Frequency histograms of the explainable variance of our neuronal samples from V1 (top) and V4 (bottom). Triangles represent the median value of each area (V1, 0.16; V4, 0.12).

Fig. 3 Prediction performance of the encoding models with Portilla-Simoncelli statistics filters

(a) The average responses of V1 and V4 neurons (black traces) to a 3-min test video across 10 trials and the predicted responses (green for V1, orange for V4). The prediction performance (r) is shown for each neuron. The scale indicates the response strength equivalent to 4 standard deviations (s.d.) of the z-scored signal strength distribution. **(b)** Frequency histograms of prediction performance for our entire database (n = 2146 neurons for V1, n = 2492 neurons for V4). Triangles represent the median values. Neurons exceeding >0.2 performance accuracy (dark columns) were further analyzed.

Fig. 4 Comparison of fitting weights between V1 and V4

(a) Fitting weights of eight groups of image statistics are plotted for each neuron examined in V1 and V4 (n = 972 for V1, n = 477 for V4). Different colors represent different image statistics groups. The brightness represents the weight amplitude; for each neuron, the greatest brightness indicates maximum weight (Max), the intermediate brightness indicates intermediate weight (more than half-maximum: >Max/2), and the black areas indicate low weight (less than half-maximum: <Max/2). **(b)** Box plots of fitting weights of the eight statistics groups. The horizontal line in each box indicates the median, the upper edge indicates the 75th percentile, and the lower edge indicates the 25th percentile. Asterisks indicate a statistically significant difference (* $p < 0.05$, ** $p < 0.01$; Wilcoxon rank-sum test, Bonferroni correction). a.u.: arbitrary unit. **(c)** Cumulative distributions of the non-spectral statistics ratio (see Materials and Methods) for V1 and V4. Triangles indicate the median values (V1, 0.24; V4, 0.34). Asterisks (***) indicate a significant difference between the two areas ($p < 10^{-5}$; Wilcoxon rank-sum test). S, spectral; M, marginal; LP, linear position correlation; LO, linear orientation correlation; LS, linear scale correlation; EP, energy position correlation; EO, energy orientation correlation; ES, energy scale correlation.

Fig. 5 Fitting weights of different parameters of marginal statistics in V1 and V4

Fitting weights are plotted for various parameters of the luminance distribution (marginal statistics), including the mean, variance (Var), skewness (Skew), Kurtosis (Kurt), Minimum (Min), Maximum (Max), sub-band Skewness (SkewL1 to SkewL4), sub-band Kurtosis (KurtL1 to KurtL4), and variance of high-pass filtered images (VarHP). See Materials and Methods for details. Asterisks (**) indicate a significant difference between the two areas ($p = 0.01$; Wilcoxon rank-sum test with Bonferroni correction for multiple comparisons for the 15 marginal statistics). a.u., arbitrary unit.

Fig. 6 Spatial distributions of neurons with different tunings to Portilla-Simoncelli statistics in V1 and V4

(a, d, g, j) Spatial maps of neurons with largest weights of different Portilla-Simoncelli statistics groups in local regions of V1 (**a**, **d**) and V4 (**g**, **j**). Neurons are color-labeled according to the statistics group to which they are most sensitive. See panel **b** for the color codes. Scales indicate 100 μm. **(b, e, h, k)** Fitting weights of eight groups of image statistics are plotted for neurons imaged at sites (**a**), (**d**), (**g**), and (**j**), respectively. See Figure 4 legend for the color codes. **(c, f, i, l)** Frequency histograms of the non-spectral

statistics ratio for sites (**a**), (**d**), (**g**), and (**j**), respectively. Triangles indicate the median values for individual sites. See Figure 4a for abbreviations of statistics categories.

Fig. 7 Proportions of neurons tuned to spectral, marginal, and correlational statistics at individual imaging sites

The percentages of neurons with the largest fitting weights for spectral, marginal, and correlation statistics are shown for individual sites in V1 (**a**, n = 17) and V4 (**b**, n = 16). The lowest rows in **a** and **b** are the sums of the fractions across all imaging sites in V1 and V4. The color code is the same as in Figure 4a.

Figures

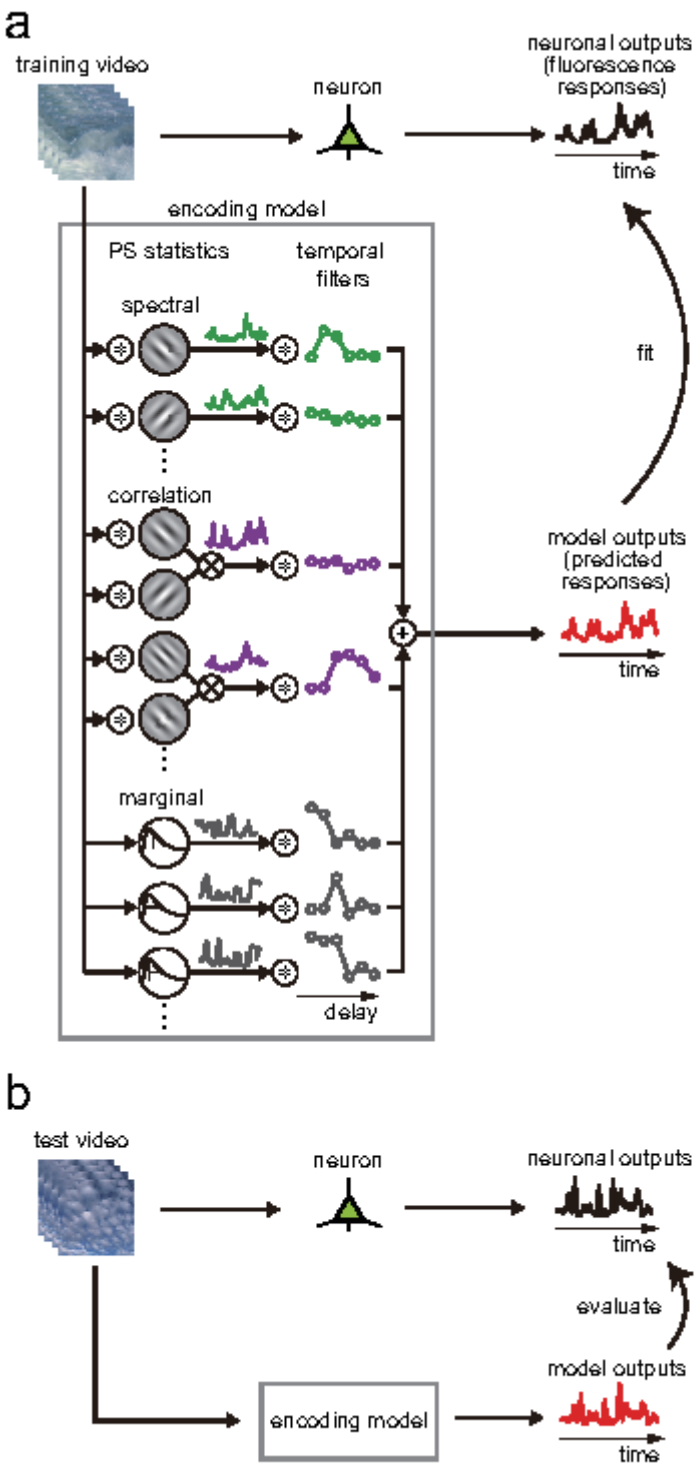

Figure 1

Encoding-model with first-stage filters computing Portilla-Simoncelli statistics (a) The first stage of the encoding model consists of a set of spectral filters, correlations between outputs of different filters, and the marginal statistics of the luminance distributions of images (PS statistics). These filters produce

time-varying outputs when a video is fed as an input. Each of these filters is followed by a temporal filter at the second stage. The model then linearly sums the second-stage filter outputs to produce the model response. For each neuron recorded, we trained the model by optimizing the temporal filters to mimic the neuron's fluorescence responses to a training video. (b) The model performance was evaluated by comparing the model response and the neuron response to a test video that was not used for the model construction. Encircled asterisks, convolution; encircled crosses, correlation; encircled plus sign, linear summation.

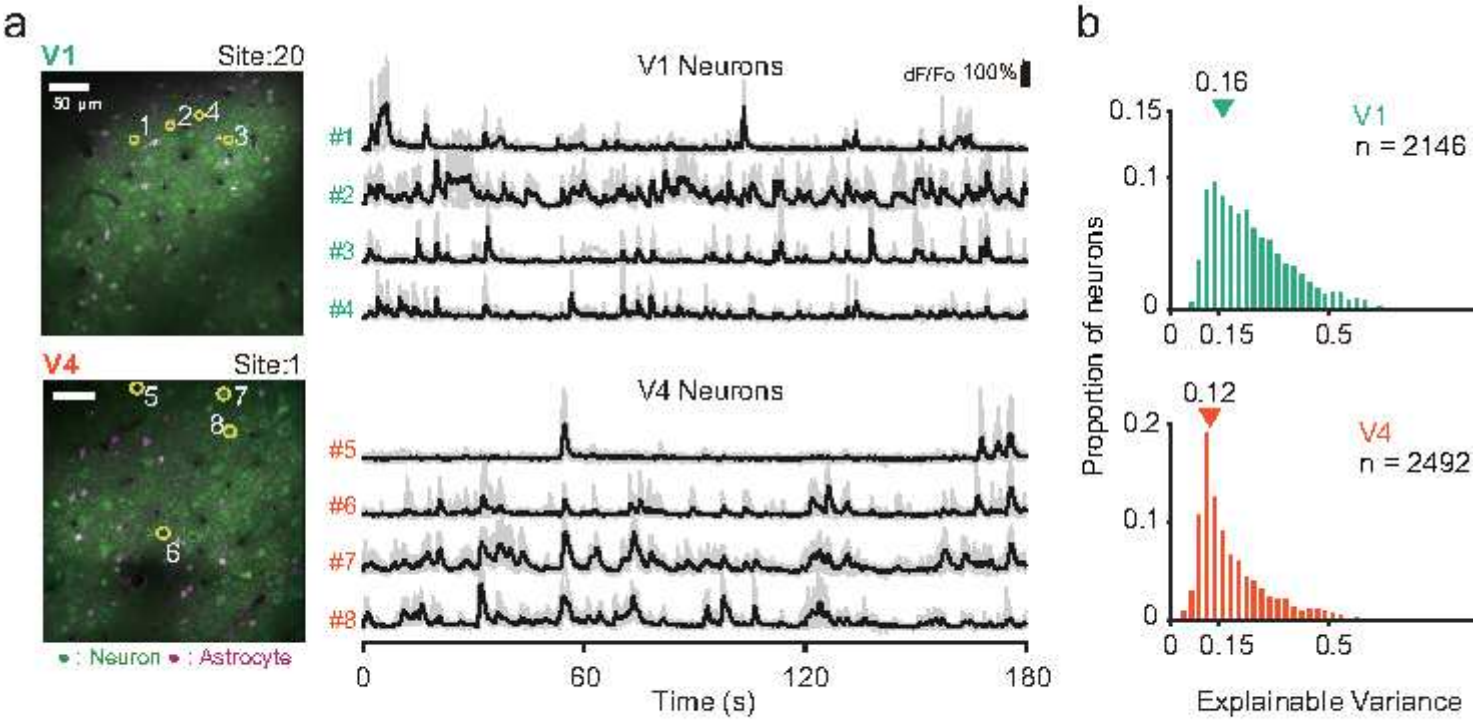

Figure 2

Reliability and diversity of fluorescence responses of V1 and V4 neurons to naturalistic videos (a) Fluorescence response time courses of example neurons in V1 (#1–#4) and V4 (#5–#8). Black lines indicate average responses across 10 trials, and gray lines indicate responses in individual trials. Numbers in the two-photon microscopy images correspond to the neuron numbers to the left of the fluorescence traces. In contrast with astroglia, neurons were sulforhodamine negative. dF/F denotes changes (dF) in fluorescence signals normalized by the DC component of the fluorescence (F0). (b) Frequency histograms of the explainable variance of our neuronal samples from V1 (top) and V4 (bottom). Triangles represent the median value of each area (V1, 0.16; V4, 0.12).

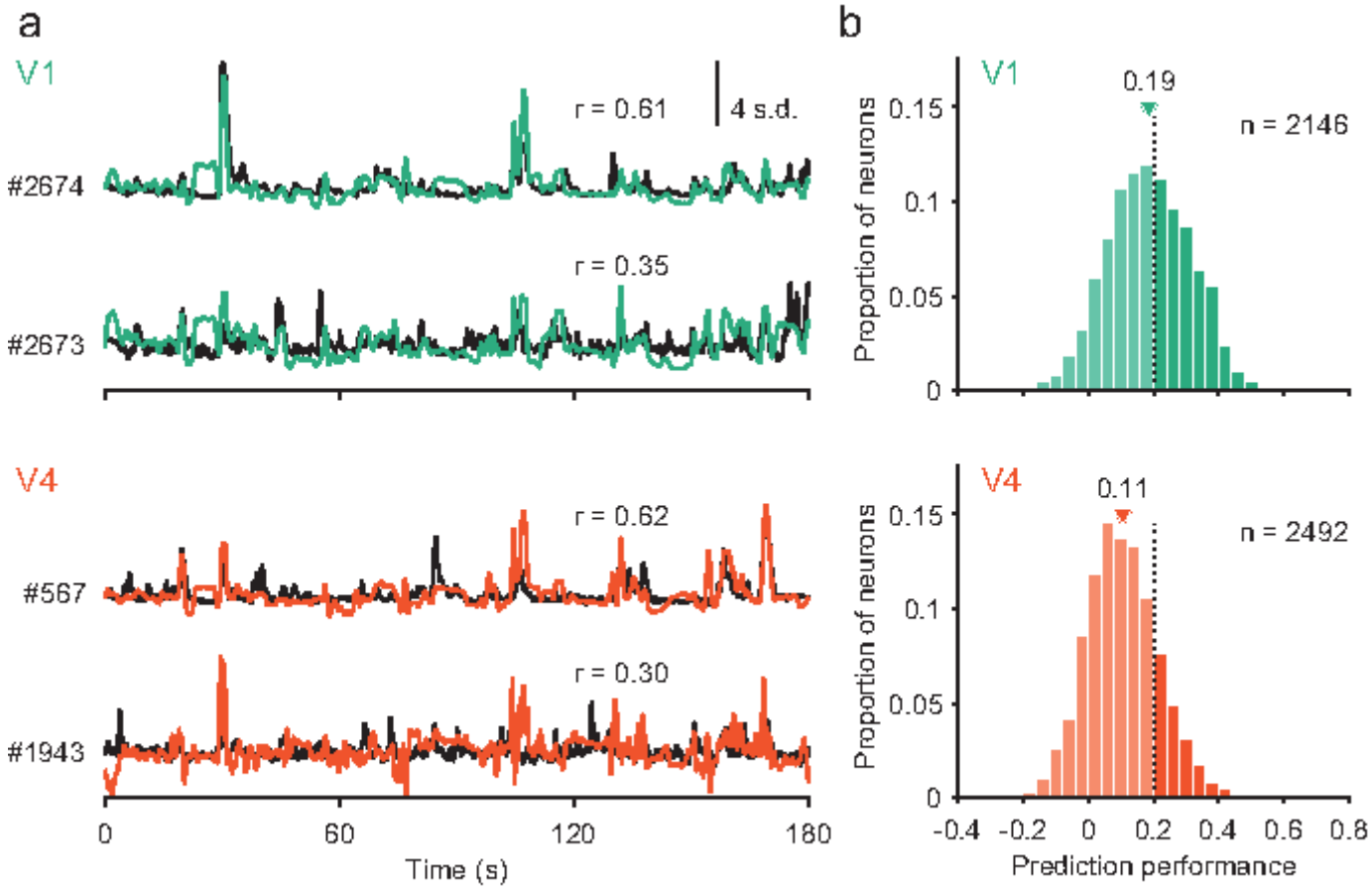

Figure 3

Prediction performance of the encoding models with Portilla-Simoncelli statistics filters (a) The average responses of V1 and V4 neurons (black traces) to a 3-min test video across 10 trials and the predicted responses (green for V1, orange for V4). The prediction performance (r) is shown for each neuron. The scale indicates the response strength equivalent to 4 standard deviations (s.d.) of the z-scored signal strength distribution. (b) Frequency histograms of prediction performance for our entire database (n = 2146 neurons for V1, n = 2492 neurons for V4). Triangles represent the median values. Neurons exceeding >0.2 performance accuracy (dark columns) were further analyzed.

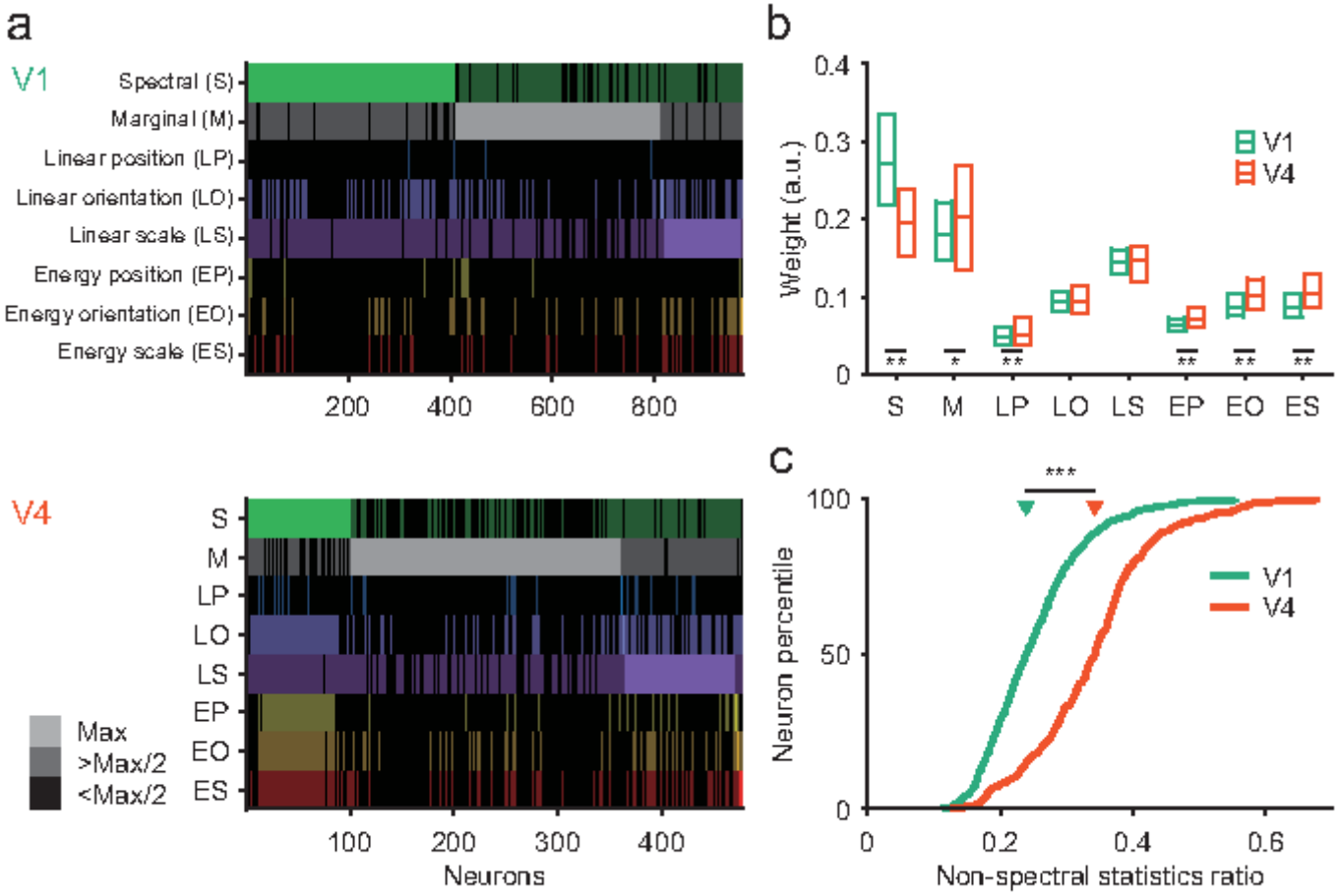

Figure 4

Comparison of fitting weights between V1 and V4 (a) Fitting weights of eight groups of image statistics are plotted for each neuron examined in V1 and V4 (n = 972 for V1, n = 477 for V4). Different colors represent different image statistics groups. The brightness represents the weight amplitude; for each neuron, the greatest brightness indicates maximum weight (Max), the intermediate brightness indicates intermediate weight (more than half-maximum: >Max/2), and the black areas indicate low weight (less than half-maximum:

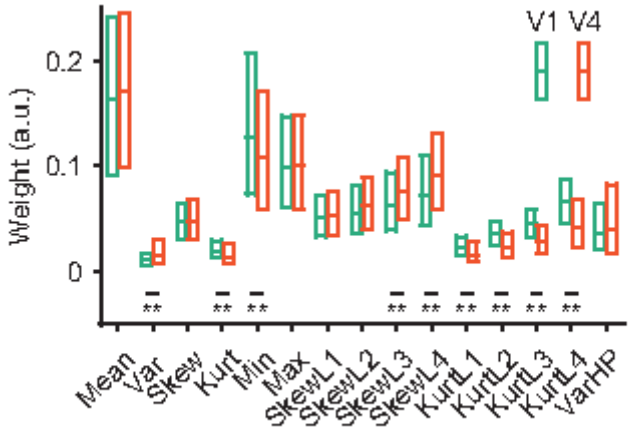

Figure 5

Fitting weights of different parameters of marginal statistics in V1 and V4 Fitting weights are plotted for various parameters of the luminance distribution (marginal statistics), including the mean, variance (Var), skewness (Skew), Kurtosis (Kurt), Minimum (Min), Maximum (Max), sub-band Skewness (SkewL1 to SkewL4), sub-band Kurtosis (KurtL1 to KurtL4), and variance of high-pass filtered images (VarHP). See Materials and Methods for details. Asterisks (**) indicate a significant difference between the two areas (p = 0.01; Wilcoxon rank-sum test with Bonferroni correction for multiple comparisons for the 15 marginal statistics). a.u., arbitrary unit.

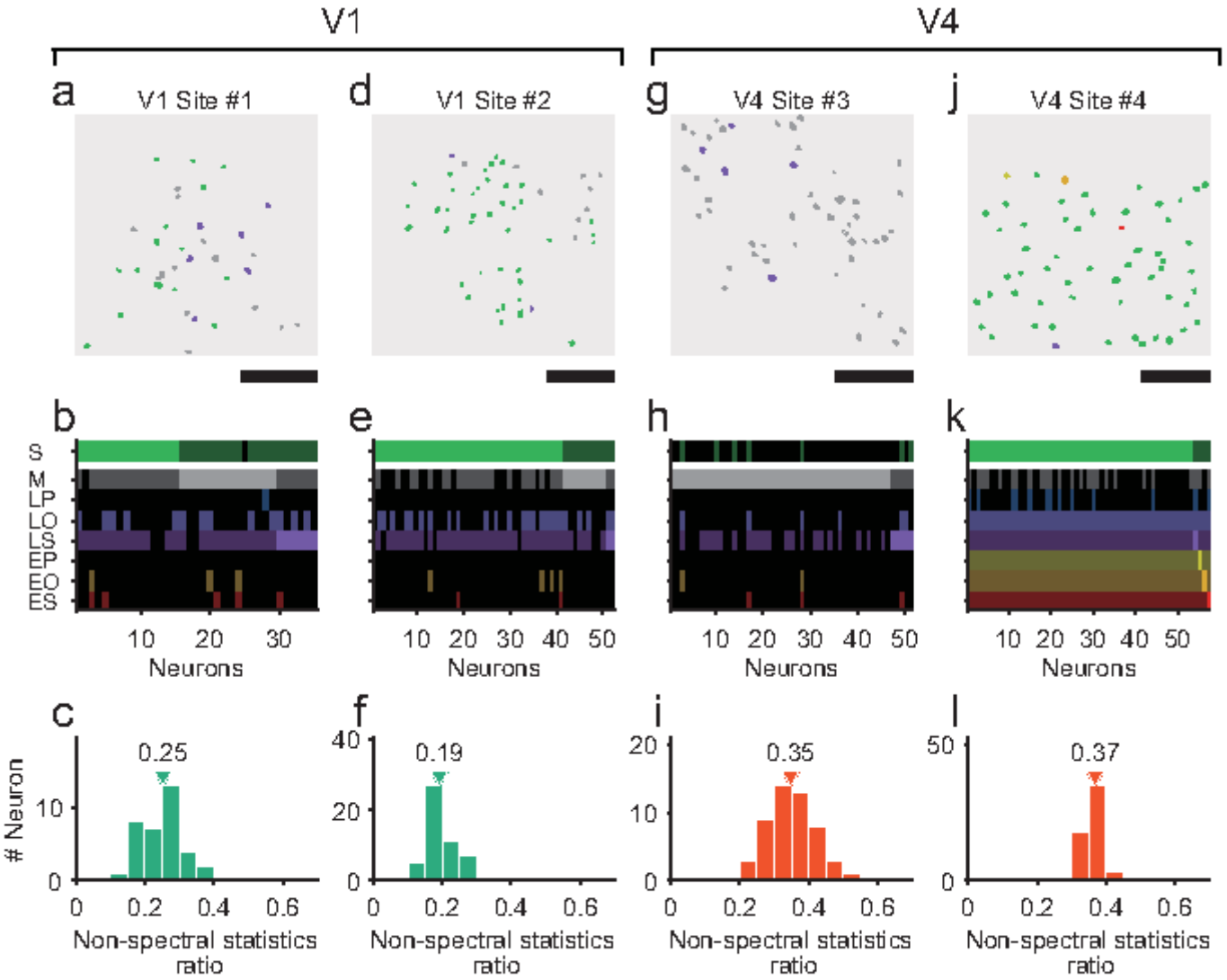

Figure 6

Spatial distributions of neurons with different tunings to Portilla-Simoncelli statistics in V1 and V4 (a, d, g, j) Spatial maps of neurons with largest weights of different Portilla-Simoncelli statistics groups in local regions of V1 (a, d) and V4 (g, j). Neurons are color-labeled according to the statistics group to which they are most sensitive. See panel b for the color codes. Scales indicate 100 μm. (b, e, h, k) Fitting weights of

eight groups of image statistics are plotted for neurons imaged at sites (a), (d), (g), and (j), respectively. See Figure 4 legend for the color codes. (c, f, i, l) Frequency histograms of the non-spectral statistics ratio for sites (a), (d), (g), and (j), respectively. Triangles indicate the median values for individual sites. See Figure 4a for abbreviations of statistics categories.

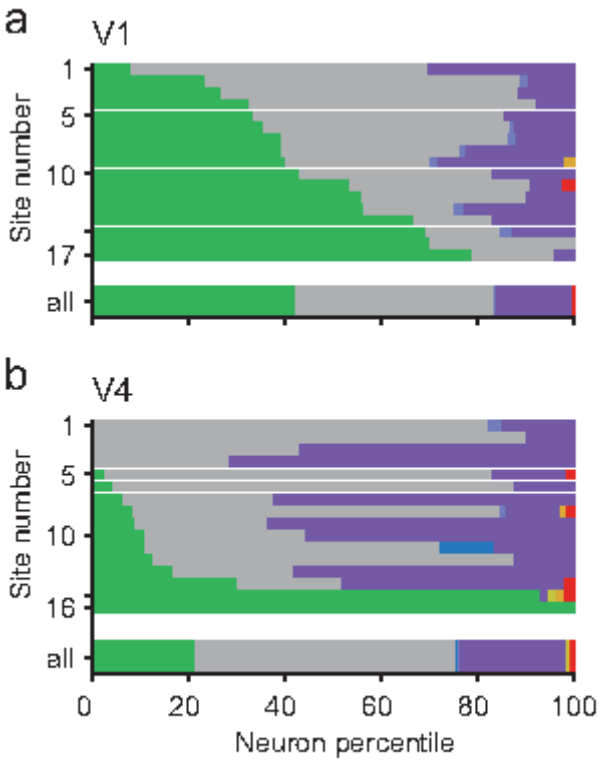

Figure 7

Proportions of neurons tuned to spectral, marginal, and correlational statistics at individual imaging sites The percentages of neurons with the largest fitting weights for spectral, marginal, and correlation statistics are shown for individual sites in V1 (a, n = 17) and V4 (b, n = 16). The lowest rows in a and b are the sums of the fractions across all imaging sites in V1 and V4. The color code is the same as in Figure 4a.