A texture statistics encoding model reveals hierarchical feature selectivity across human visual cortex

Margaret M. Henderson[1,2,3], Michael J. Tarr[1,2,3], and Leila Wehbe[1,2,3]

[1]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, USA
[2]Department of Psychology, Carnegie Mellon University, Pittsburgh, USA
[3]Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA

September 24, 2022

Abstract

Mid-level visual features, such as contour and texture, provide a computational link between low- and high-level visual representations. While the detailed nature of mid-level representations in the brain is not yet fully understood, past work has suggested that a texture statistics model (P-S model; Portilla Simoncelli, 2000) is a candidate for predicting neural responses in areas V1-V4 as well as human behavioral data. However, it is not currently known how well this model accounts for the responses of higher visual cortex regions to natural scene images. To examine this, we constructed single voxel encoding models based on P-S statistics and fit the models to human fMRI data from the Natural Scenes Dataset (Allen et al., 2021). We demonstrate that the texture statistics encoding model can predict the held-out responses of individual voxels in early retinotopic areas as well as higher-level category-selective areas. The ability of the model to reliably predict signal in higher visual cortex voxels suggests that the representation of texture statistics features is widespread throughout the brain, potentially playing a role in higher-order processes like object recognition. Furthermore, we use variance partitioning analyses to identify which features are most uniquely predictive of brain responses, and show that the contributions of higher-order texture features to model accuracy increases from early areas to higher areas on the ventral and lateral surface of the brain. These results provide a key step forward in characterizing how mid-level feature representations emerge hierarchically across the visual system.

Introduction

Information in visual cortex is processed by a series of hierarchically organized brain regions, with the complexity of representations increasing at each level. While there are intuitive explanations for response properties at the ends of this hierarchy (e.g., oriented spatial frequency filters in primary visual cortex (Carandini et al., 2005; Hubel & Wiesel, 1962) or object and category representations in inferotemporal cortex (Desimone et al., 1984; Grill-Spector & Weiner, 2014), the computations performed at intermediate levels have proven more challenging to describe. These intermediate- or mid-level visual areas are thought to represent features like contour and texture, which play an important role in figure-ground segmentation, shape processing, and object and scene classification (Bergen & Landy, 1991; Connor et al., 2007; Peirce, 2015; Ullman et al., 2002; Walther & Shen, 2014). Thus, developing a robust model of mid-level representation is fundamental for understanding how the visual system extracts meaningful information from the environment.

Computational texture models, particularly the influential model proposed by Portilla and Simoncelli (Portilla and Simoncelli, 2000; hereafter referred to as the "P-S model"), have proven useful in understanding the mid-level features that drive visual responses. The P-S model is constructed using a steerable pyramid decomposition (Simoncelli & Freeman, 1995) to extract features at different spatial scales and orientations, and consists of features at multiple levels of complexity, including the linear and energy responses of each pyramid sub-band, which approximate V1 simple and complex cells, respectively (Figure 1, left panels). The model also includes higher-order correlations among the pyramid sub-bands, which provide information about mid-level image structure. The P-S model can account for aspects of texture sensitivity in primate visual areas such as V1, V2, and V4 (Freeman et al., 2013; Hatanaka et al., 2022; Okazawa et al., 2017, 2015), with sensitivity to the higher-order model features increasing from V1 to V2 to V4 (Freeman et al., 2013; Hatanaka et al., 2022; Okazawa et al., 2017). In humans, functional magnetic resonance imaging (fMRI) studies suggest that P-S statistics can account for some response properties in early visual and ventral visual cortex (Baumgartner & Gegenfurtner, 2016; Long et al., 2016). The P-S model also captures key aspects of human behavior, including

perceptual discrimination of texture patches (Balas, 2006; Freeman & Simoncelli, 2011), crowding (Balas et al., 2009), search performance (Rosenholtz et al., 2012) and discrimination of high-level object properties such as real-world size and animacy (Long et al., 2016; Long et al., 2017).

Despite these promising results, it is not yet known how effective the P-S model is at predicting neural responses to natural scene images, particularly in higher visual cortex. The majority of studies using the P-S model have focused on synthetic stimuli or textures that are relatively homogeneous across space (e.g., Freeman and Simoncelli, 2011; Okazawa et al., 2017, 2015). In contrast, natural scenes, which include objects and other complex, spatially localized elements, may prove more challenging to characterize (Portilla & Simoncelli, 2000). One exception is a two-photon imaging study that examined the P-S model's prediction of neural responses in macaque V1 and V4 for naturalistic stimuli (Hatanaka et al., 2022); however, this study did not consider higher visual cortex. Under the hypothesis that mid-level features also play a key role in the neural processing of object and scene categories (Bracci et al., 2017; Connor et al., 2007; Groen et al., 2017; Long et al., 2018; Nasr et al., 2014), modeling responses to natural images within higher visual cortex is a critical test of the P-S model's ability to generalize to real-world vision.

To address this question, we developed a forward encoding model (Naselaris & Kay, 2015; Serences & Saproo, 2012) that uses P-S model texture statistics to predict voxelwise responses from the Natural Scenes Dataset (NSD; Allen et al., 2021). Our model incorporates parameters that independently describe the feature selectivity and spatial selectivity of individual voxels (adapted from St-Yves and Naselaris, 2017). An explicit spatial receptive field for each voxel improves upon past work (Hatanaka et al., 2022; Okazawa et al., 2015), allowing us to adaptively fit voxel responses with a wide range of receptive field positions and sizes. Voxelwise spatial receptive fields also aid model interpretability, allowing examination of the spatial and feature selectivity of individual voxels in the fitted model. To probe the contributions of specific feature classes, we used variance partitioning to determine how much of the explained variance was uniquely attributable to each subset of texture model features.

Critically, our model: (1) achieves accurate predictions of held-out voxel responses throughout the visual hierarchy, with highest performance in early and mid-level areas; (2) recovers the hierarchical organization of visual cortex, with retinotopic areas best described by lower-level features and anterior category-selective areas best described by higher-level features. These and other results we present here facilitate a better understanding of how hierarchical computations give rise to intermediate-level and high-level representations in the brain.

Methods

Acquisition and preprocessing of fMRI data

We utilized the Natural Scenes Dataset (NSD), a large-scale, publicly available fMRI dataset. Full details on the acquisition of this data can be found in Allen et al., 2021. The NSD includes whole-brain BOLD fMRI measurements from 8 human subjects who each viewed between 9,000-10,000 natural scene images over the course of a year (between 30-40 scan sessions). Functional scanning was conducted at 7T using whole-brain gradient-echo EPI at 1.8 mm resolution and 1.6 second repetition time (TR). Images were sampled from the Microsoft Common Objects in Context (COCO) database (Lin et al., 2014). Over the course of the experiment, each image was viewed approximately 3x, for a total of roughly 30,000 trials per subject. Of the approximately 10,000 images viewed by each subject, ∼9,000 images were seen only by that subject, and 907 were viewed at least once by each subject. Each image was presented in color, at a size of 8.4° x 8.4° (° = degrees of visual angle), and was viewed for a duration of 3 seconds, with 1 second between trials. Throughout each scan, subjects performed a task in which they reported whether each image was new or old (i.e., whether or not it had been presented before in any session), while fixating centrally on a small fixation dot superimposed on each image.

As described in Allen et al., 2021, the functional data from all sessions were pre-processed using temporal interpolation to correct for slice timing differences and spatial interpolation to correct for head motion (see Allen et al., 2021 for details). All analyses were performed in each subject's native volumetric space (1.8 mm resolution voxels). A general linear model was used to estimate beta weights for each voxel and each individual trial (Prince et al., 2022). We obtained the beta weights from Allen et al., 2021 after this stage, and we then performed a few additional steps to prepare the beta weights for our analyses. First, beta weights for each voxel were z-scored across all trials of each scan session. To improve the signal-to-noise ratio of the data, we then averaged the beta weights for each voxel across trials where the same image was shown (approximately three trials/image), resulting in a single value for each voxel in response to each of the unique (∼10,000) images. Note that for subjects who did not complete the entire experiment, there were fewer than 10,000 images.

We masked out a broad region of interest in visual cortex to include in all analyses. This region included voxels that were part of the "nsdgeneral" ROI described in Allen et al., 2021, which is meant to capture the general spatial extent of voxels that were responsive to the NSD image stimuli. To broaden the scope of brain areas included in our analyses, we additionally included voxels belonging to any ROI within several sets of ROI definitions: any ROI in the probabilistic atlas provided in Wang et al., 2015, which includes regions of the intraparietal sulcus, any ROI belonging to an early retinotopic area based on pRF mapping, and any face-selective, body-selective, or place-selective ROI identified through a functional category localizer task (see *Methods: Defining regions of interest (ROIs)* for details on pRF and category localizer tasks). This large mask defines the extent of voxels that are included in whole-brain surface maps (e.g., Figure 2A). We also computed summary statistics at the ROI level using more fine-grained definitions of individual ROIs (e.g., Figure 2B), see *Methods: Defining regions of interest (ROIs)* for details.

Defining regions of interest (ROIs)

We defined retinotopic and category-selective ROIs based on functional localizers that were collected as part of the NSD experiment. A category localizer task (Stigliani et al., 2015) was used to define place-selective regions (parahippocampal place area/PPA, occipital place area/OPA, and retrosplenial cortex/RSC), a face-selective region (fusiform face area/FFA; we combined FFA-1 and FFA-2 into a single FFA region), and a body-selective region (extrastriate body area/EBA). A pRF mapping task (sweeping bar stimuli; Benson et al., 2018) was used to define early retinotopic visual ROIs V1, V2, V3, and hV4 (see Allen et al., 2021 for more details). When retinotopic and category-selective ROIs were overlapping with one another, we excluded any voxels that were overlapping from the retinotopic definition and added them to the corresponding category-selective ROI only. To ensure that the final set of category-selective ROIs were non-overlapping, we always prioritized face-selective ROIs over place- and body-selective ROIs, and prioritized place-selective ROIs over body selective ROIs.

Noise ceiling estimation

The noise ceiling is a measure that estimates the proportion of variance in each voxel's response that can theoretically be explained by properties of the stimulus. Our measure of noise ceiling was computed by Allen et al., 2021, by estimating the variability of each voxel's response across presentations of the same image (method described in Allen et al., 2021). Note that since we averaged over repeated presentations of each image, this makes the noise ceiling higher than for single trial data. We used the noise ceiling as an additional threshold to restrict the voxels used our whole-brain and ROI-level analyses, using a threshold of 0.01.

Texture statistics encoding model

Overview

Our encoding model incorporated parameters for each voxel's feature selectivity as well as its spatial selectivity (St-Yves & Naselaris, 2017), see Figure 1 for an illustration. To fit the model, we first created a grid of candidate population receptive fields or pRFs (see *Methods: Population receptive fields (pRFs)*), and extracted texture statistics features (see *Methods: Texture statistics features*) within each candidate pRF. For each pRF, we then used regularized regression to fit a linear model that predicts the voxel's activation as a function of the texture statistics features corresponding to that pRF. The final encoding model for each voxel was computed by identifying the best-fitting model over all candidate pRFs. More detail on each step is provided in the following sections.

Population receptive fields (pRFs)

We modeled each candidate population receptive field (pRF) as a 2-dimensional Gaussian (Dumoulin & Wandell, 2008; St-Yves & Naselaris, 2017). Each pRF has three parameters, x_0, y_0, and σ, where $[x_0, y_0]$ and σ, respectively, indicate the center and standard deviation of the 2-dimensional Gaussian response profile:

$$g_{x_0,y_0,\sigma}(x,y) = \exp\left(-\left(\frac{(x-x_0)^2}{2\sigma^2} + \frac{(y-y_0)^2}{2\sigma^2}\right)\right) \tag{1}$$

Our grid of candidate pRF parameters was designed to approximate a log-polar grid, where the spacing between adjacent pRF centers is linear in terms of polar angle position (θ), and nonlinear in terms of eccentricity (r), such that candidate centers were more closely spaced closer to the center of the visual field. This nonlinear eccentricity spacing was intended to account for the cortical magnification factor in human visual cortex, where the neuronal sampling of visual space is more dense close to the fovea (Duncan & Boynton, 2003). Specifically, we used 16 candidate polar angle positions that were linearly spaced, ranging from 0° to 337.5° in steps of

22.5°, and 10 candidate eccentricities that were logarithmically spaced, ranging from 0° to 7°. For each of these candidate centers, we generated 10 candidate σ values that were also spaced logarithmically, ranging from 0.17° to 8.4°. The complete grid of pRF parameters over all combinations of r, θ, and σ resulted in 1,600 pRFs.

Since our eccentricity range extended slightly beyond the physical image extent (an 8.4° square), some of the smaller pRFs at the largest r values were entirely non-overlapping with the image region (in contrast, larger pRFs at these large eccentricities were still partially overlapping with the image region). To address this issue, we removed from the grid any pRFs whose rough spatial extent (center$\pm\sigma$) was non-overlapping with the image region. This resulted in 1,456 pRFs in the final grid.

Texture statistics features

As the first step of our voxelwise encoding model framework, we extracted a set of image-computable texture features which are meant to capture various aspects of local image structure. We utilized the aforementioned P-S model (Portilla & Simoncelli, 2000), which has previously been shown to predict both human behavioral judgments of textures and neural responses in mid-level visual regions such as V2 and V4 (Freeman & Simoncelli, 2011; Freeman et al., 2013; Okazawa et al., 2015; Portilla & Simoncelli, 2000). Full details of the model construction and its motivation are given in (Portilla & Simoncelli, 2000) and (Freeman & Simoncelli, 2011). Here, we provide a brief description of the model components and describe how we have incorporated it into our pRF modeling framework.

The first stage of the model consists of using a steerable pyramid (Simoncelli & Freeman, 1995) to decompose each image into a set of orientation and frequency sub-bands (we used 4 orientations and 4 frequency bands). This step was implemented using the Python package *pyrtools* to construct a steerable pyramid in the frequency domain, and was performed on grayscale images at a resolution of 240 x 240 pixels. The steerable pyramid results in spatial maps of complex-valued coefficients at each scale and orientation, from which we can compute the real part and the magnitude, which correspond approximately to the responses of V1 simple and complex cells, respectively (Freeman & Simoncelli, 2011). The four frequency bands output by our pyramid were centered at approximately 0.9, 1.8, 3.6, and 7.1 cycles per degree (cpd), and the four orientation bands were centered at 0°, 45°, 90° and 135°. The steerable pyramid additionally computes high-pass and low-pass residual images, as well as a partially-reconstructed low-pass image representation at each scale.

Our complete texture statistics model included 10 sub-types of features, resulting in 641 total features, each of which we computed at each pRF grid position. For simplicity, we have divided the features into two sub-groups for several of our analyses (e.g., Figure 3). The first sub-group, which we have termed "lower-level texture features", consists of marginal statistics computed from either the raw image luminance values, or from the outputs of the steerable pyramid. The second set of features, which we have termed "higher-level texture features", consists of higher-order correlations computed from the steerable pyramid, generated by either correlating different orientation/scale channels of the pyramid, or spatially shifted versions of the same channel of the pyramid. When computing these features, we used the Gaussian profile for the pRF of interest (Equation (1)) as a weighting matrix (similar to the "pooling region" used in Freeman and Simoncelli, 2011). All feature extraction steps after the initial steerable pyramid computation were done using custom code in PyTorch.

The first set of lower-level features, termed *pixel* features, consists of the pRF-weighted mean, variance, skew, and kurtosis of the raw pixel luminance values. The second and third sets of features, *energy-mean* and *linear-mean* features, are the pRF-weighted mean of the magnitude and real part, respectively, of each steerable pyramid feature channel (4 orientations x 4 scales). The fourth set of lower-level features, *marginal* features, includes the pRF-weighted skew and kurtosis of the low-pass pyramid reconstruction at each scale, and the pRF-weighted variance of the high-pass residual image.

The first set of higher-level texture features consists of autocorrelations of the steerable pyramid features, which are computed by correlating each feature map with spatially-shifted versions of itself. Autocorrelations were computed from the magnitude of each main pyramid feature channel (*energy-auto*), the low-pass reconstruction at each scale, and the high-pass residual image. The autocorrelations of the low-pass reconstructions and the high-pass residual image were treated as a single combined group of features for the variance partitioning analysis (*linear-auto*), but were treated separately when performing PCA (see below), as this yielded higher overall model accuracy. To compute each autocorrelation matrix, we cropped out a square region of the image that spanned approximately $\pm 2\sigma$ from the pRF center. We then computed the weighted 2-dimensional autocorrelation over this cropped image region, using the pRF profile (cropped in the same way) as a weighting matrix. From this matrix, we retained a fixed number of pixels from the center, so that the number of autocorrelation features returned was independent of pRF size.

The remaining higher-level texture features consist of cross-correlations, computed by correlating different feature maps output by the steerable pyramid. All cross-correlations were computed using the entire image, weighted by the pRF profile. The first two sub-sets of cross-correlation features (*energy-cross-orient*) and *linear-cross-orient*) are cross-correlations of feature maps (either magnitudes or real parts) at the same scale, but with different orientations. The next sub-sets of features (*energy-cross-scale*) were computed by correlating the magnitude of feature maps having the same orientation but different scales, after up-sampling the resolution of the map at the coarser scale and doubling its phase. An additional sub-set of features (*linear-cross-scale*) were computed similarly, but using the real or imaginary component of the feature maps only. An additional two sub-sets of features were computed by correlating the low-pass residual image with spatially shifted versions of itself (within-scale) or with the lowest frequency pyramid feature map (cross-scale). The within-scale group of these features were included in the *linear-cross-orient* group for the variance partition analysis, while the cross-scale group were included in the *linear-cross-scale* group for the variance partition analysis, but these groups were treated separately when performing PCA (see next paragraph), as this grouping tended to result in better model performance.

To reduce the dimensionality of the texture model feature space and prevent over-fitting, we performed principal components analysis (PCA) on the higher-level texture features before using them to construct encoding models. PCA was performed within each pRF separately, within each sub-set of higher-level texture features individually (note that some feature groups were further subdivided when performing PCA, see above paragraphs). No dimensionality reduction was performed on the lower-level texture features. Where PCA was used, the principal components were always computed using the training data only, and all data including the validation data was projected into the same subspace. We retained the minimum number of components necessary to explain 95% of the variance. Since PCA was performed on the features from one pRF at a time, this meant that the dimensionality of the features was not required to be the same across all pRFs.

Model fitting procedure

To construct the texture statistics encoding model, we modeled each voxel response as a weighted sum of the texture statistics features corresponding to each image and each pRF (plus an intercept). As described in the previous section (*Methods: Texture statistics features*), texture statistics features were computed in a spatially specific manner, such that the feature activations for a given image depend on the pRF parameters x_0, y_0 and σ. We solved for the weights of the encoding model for each voxel using ridge regression (L2-regularization), as used in previous work (Güçlü & van Gerven, 2014; Huth et al., 2016; Wehbe et al., 2014). To improve the regression fits, we used a banded ridge regression method (Nunez-Elizalde et al., 2019) which incorporated two different ridge regularization (λ) parameters, one corresponding to the lower-level texture features and one corresponding to the higher-level texture features. Both λ parameters were selected on a per-voxel basis from a set of 10 candidate λ values logarithmically spaced between 0 - 10^5. To sample all combinations of two λ values, we created a grid of all 100 possible λ combinations. Cross-validation was used to determine the best λ parameters for each voxel; the full cross-validation procedure is as follows (see also Figure 1).

First, we held out a validation set of ∼1,000 images from the ∼10,000 total images for each subject. This set of images always consisted of the shared images in which every subject saw the same images (note that for subjects who did not complete the entire NSD experiment, there were fewer images in this validation set; minimum was 907 images; see *Methods: Acquisition and preprocessing of fMRI data*). The remaining ∼9,000 images made up the training data. Following this, we held out a random 10% of the training data as a nested validation set. This nested set was used to select the ridge parameter, as well as the best pRF parameters for each voxel. Using the remaining 90% of the training data, we then estimated regression weights for each of our candidate λ values, as well as for each of our candidate pRF models. Based on the estimated weights, we computed the loss for each lambda and each pRF, by generating a prediction of the nested validation data and computing the sum of the squared error for this prediction. We then selected the best pRF parameters and λ values for each voxel based on which values resulted in the lowest loss. The resulting pRF parameters and regression weights made up the final encoding model for the voxel. Finally, to estimate overall model accuracy, we generated predicted responses of each voxel on the held-out validation set and computed the coefficient of determination (R^2) between the actual and predicted response. To make plots of R^2 in surface space for each subject (i.e., Figure 2A), we utilized PyCortex (Gao et al., 2015).

Permutation testing

For each voxel, we used a permutation test to evaluate whether the texture statistics encoding model resulted in higher than chance accuracy at predicting the validation set data. To perform the permutation test, we randomly shuffled (1,000 times) the image labels corresponding to each voxel activation pattern (voxel activations were already averaged over presentations of each repeated image before this shuffling was performed; see *Methods: Acquisition and preprocessing of fMRI data*). Shuffling was always performed within the training set, validation set, and nested validation set (see *Methods: Model fitting procedure*) separately. For each shuffling iteration, we performed the regression procedure from scratch, including fitting the model weights and computing R^2. However we did not re-fit each voxel's pRF on the shuffled data, instead we used the best pRF for each voxel as determined from the intact data, and only re-fit the regression weights. To compute a one-tailed p-value for each individual voxel, we calculated the number of iterations on which the shuffled R^2 was $\geq$ the real R^2, divided by the number of iterations. We then performed FDR correction on the p-values for all voxels from each subject (q=0.01; Benjamini and Hochberg, 1995). We used the result of this significance test as a mask to determine which voxels to include in the subsequent variance partitioning analyses.

Variance partitioning

In order to determine the unique contribution of different subsets of texture features to encoding model predictive accuracy, we used a variance partitioning analysis (see Groen et al., 2012; Lescroart and Gallant, 2019 for similar approaches). When performing this analysis, we always restricted the set of voxels to those that had above-chance accuracy for the full texture statistics encoding model (see previous section). The overall approach was as follows. First, we fit the full texture statistics encoding model with all feature sets concatenated, using the method described in the previous section. We computed the prediction accuracy of the concatenated model, R^2_{full}. Then, we progressively removed one set of features at a time from the full model, and fit the weights of each partial model. When constructing the partial models, we always used the features corresponding to the voxel's best pRF (as determined based on fitting the full concatenated feature space). This ensured that the differences between models were due solely to differences in features and not to changes in the estimated pRF parameters. Then, for each partial model, we generated a prediction of the validation set data and and computed R^2 for the partial model, $R^2_{partial}$. Finally the unique variance attributable to each feature subset was computed as:

$$R^2_{unique} = R^2_{full} - R^2_{partial} \tag{2}$$

Where $R^2_{partial}$ denotes the partial model where the feature set of interest was removed. Occasionally, this procedure can result in values where $R^2_{unique} < 0$, for example in cases where $R^2_{full} < R^2_{partial}$, which can happen sometimes since R^2 is evaluated on a held-out data partition and over-fitting is possible, particularly in noisy voxels where R^2_{full} is close to zero. To correct for such values in our analysis, we set all negative values for R^2_{unique} to 0.

We performed two versions of the variance partitioning analysis: in the first version, we grouped all of the lower-level or all of the higher-level texture features together (Figure 3); for the second version, we analyzed each of the 10 feature subsets individually (Figure 5, Figure 6). For the first analysis, we additionally report the variance that is shared between the lower- and higher-level feature subsets. The shared variance between two feature subsets A and B, was computed as:

$$R^2_{sharedAB} = R^2_{uniqueA} + R^2_{uniqueB} - R^2_{full} \tag{3}$$

To evaluate the significance of the unique variance explained by each feature subset, we used a bootstrapping analysis. To make this analysis computationally feasible, we performed bootstrapping only on the validation set data (i.e., when computing R^2 but not when fitting model weights). We performed 1,000 iterations of the bootstrapping test; on each iteration we resampled with replacement n images from the total n validation set images. We used the same resampling order for the image labels and for the voxel data, so that the correspondence between images and voxel responses was intact, but the exact set of images included differed on each bootstrap iteration. Using the same resampling order, we computed R^2 for the full model and R^2 for each partial model, then used these to compute R^2_{unique} for each feature set, as described above. This resulted in a distribution of 1,000 values for R^2_{unique}. We then used this distribution to compute a p-value for whether R^2_{unique} was significantly higher than zero, by computing the number of iterations on which $R^2_{unique} \leq 0$ and dividing by the number of iterations. The p-values for each subject were FDR corrected across all voxels (q=0.01; Benjamini and Hochberg, 1995).

We additionally computed whether the unique variance for each feature set was significant at the ROI-averaged level for each subject. This was done by averaging the bootstrap distributions for R^2_{unique} across all voxels in each ROI (always averaging values for the same bootstrapping iteration together), and then computing a one-tailed p-value using the same method described above. This resulted in a single value for each ROI, each subject, and each feature set. We then performed multiple comparisons correction across all values using the Holm-Bonferroni method (q=0.01; Holm, 1979). We used the Holm-Bonferroni method because the FDR procedure is not appropriate for controlling for a relatively small number of comparisons. The number of subjects with significant unique variance explained for each feature set is reported in Table 3-1 and Table 6-1. To test whether the average unique variance across all subjects was significant, we used the same averaging method just described but additionally averaged the bootstrap distributions over subjects before computing p-values, and corrected for multiple comparisons as described above.

To compute whether the unique variance explained by the lower- and higher-level texture features was significantly different for each ROI, we used the ROI-averaged bootstrap distributions to compute the distribution of differences between the lower-level unique variance and the higher-level unique variance. We then computed a two-tailed p-value by computing the proportion of iterations for which the difference was positive and the proportion for which the difference was negative, taking the minimum and multiplying by 2. We again performed multiple comparisons correction using the Holm-Bonferroni method (q=0.01).

Fitting with fixed pRF size parameter

To dissociate the effects of pRF size and feature selectivity, we performed an additional analyis in which the pRF size parameter (σ) was fixed at a single value for all voxels. To achieve this, we performed our entire model fitting pipeline (see *Methods: Model fitting procedure*) with a restricted grid of candidate pRFs. This restricted pRF grid consisted of all the pRFs in our main grid (see *Methods: Population receptive fields (pRFs)*) having the σ value of interest (thus, the voxels could be fit with different pRF centers, but had to have the same pRF size). We performed this entire procedure for each of the 10 σ values in our original pRF grid.

Results

To model neural selectivity for mid-level image structure, we constructed a forward encoding model based on image-computable texture statistics features (Figure 1; see also *Methods: Texture statistics encoding model*). The texture statistics encoding model includes parameters that capture the spatial selectivity of voxels (population receptive fields or pRFs, Dumoulin and Wandell, 2008) as well as their selectivity for a range of features related to local image structure (computed using P-S statistics; Portilla and Simoncelli, 2000). The modeled features include low-level properties of the image, including the activation in different orientation and frequency channels as extracted by a steerable pyramid (Simoncelli & Freeman, 1995), and the marginal statistics of the pixel luminances; we refer to these collectively as "lower-level" texture features (see *Methods: Texture statistics features* for more details). The model also includes features that capture higher-order structure in the image, including the cross-correlations between different orientation and frequency maps output by the steerable pyramid, as well as auto-correlations computed from individual steerable pyramid maps; we refer to these as "higher-level" texture features. All parameters of the model were fit on a single voxel basis, and overall accuracy for each voxel was quantified by generating predicted voxel responses to a set of held-out images that were not used during model training and computing the coefficient of determination (R^2) for each voxel. To facilitate comparisons across visual regions with different functional roles in mid-level visual processing, we fit and evaluated the model across a large portion of occipotemporal cortex, and also computed summary statistics for several functional ROIs: early retinotopic areas V1-hV4, scene-selective areas OPA, PPA, RSC, face-selective area FFA, and body-selective area EBA (see *Methods: Defining regions of interest (ROIs)* for details).

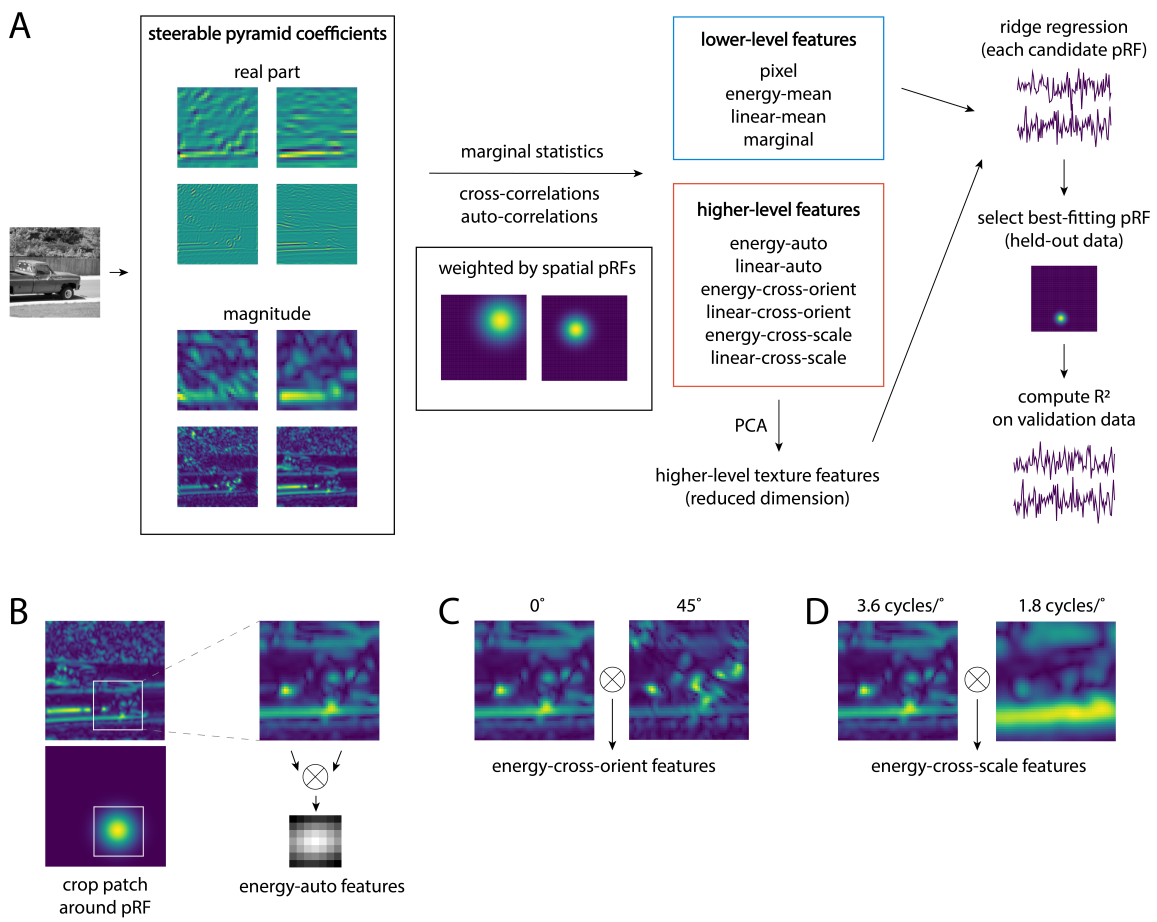

Figure 1: **(A)** Overview of the texture statistics model (Portilla & Simoncelli, 2000) and voxelwise encoding model fitting procedure. For each image, we first use a steerable pyramid (Simoncelli & Freeman, 1995) to decompose the image into four orientation and four frequency bands (note that only 2 orientations and frequencies are shown here for illustration). We then use the steerable pyramid representation to extract lower-level texture features, which include marginal statistics of the pyramid coefficients, and higher-level texture features, which consist of higher-order correlations of the pyramid coefficients (see *Methods: Texture statistics features* for details). Each of these features is computed for each population receptive field (pRF) in a grid of candidate pRFs, using the pRF as a spatial weighting matrix. Before fitting the model, we reduce the dimensionality of the higher-level texture features using PCA. We then use ridge regression to fit a set of weights for each voxel that predict its response to each image as a weighted sum of all the texture statistics features extracted from that image. We perform this fitting separately for each candidate pRF, and use the loss on a held-out data subset to choose the best pRF. Finally, we compute the model's accuracy (R^2) on a held-out validation set. See *Methods: Model fitting procedure* for more details. **(B-D)** Illustrations of how the higher-level model features are computed based on coefficients of the steerable pyramid. **(B)** The *energy-auto* features are computed by correlating each magnitude feature map with spatially-shifted versions of itself (i.e., autocorrelations). **(C)** The *energy-cross-orient* features are computed by cross-correlating magnitude feature maps corresponding to the same scale but different orientation. **(D)** The *energy-cross-scale* features are computed by cross-correlating magnitude feature maps corresponding to different scales, after up-sampling the map corresponding to the coarser scale. The *linear-auto*, *linear-cross-orient*, and *linear-cross-scale* features are computed similarly, but using the real parts of steerable pyramid coefficients.

The texture statistics encoding model achieved high predictive accuracy (R^2) across multiple visual areas, with particularly strong performance in early visual cortex (Figure 2 and Figure 2-1). Though performance progressively declined from V1 through more anterior category-selective visual regions, performance remained moderately high in voxels throughout the visual hierarchy, with the majority of voxels in all visual ROIs having above-chance validation set accuracy (Figure 2C; one-tailed p-values obtained using permutation test, corrected for multiple comparisons; q=0.01). In all areas, model performance generally scaled with the noise ceiling of individual voxels (Figure 2D). In addition, examining the spatial fit parameters of the model (Figure 4-1A) demonstrates that the model recovers single-voxel pRFs with properties that are consistent with past work (Dumoulin & Wandell, 2008), namely the tendency of pRF size (σ) to increase from early to higher-level visual

ROIs. These results provide validation of our modeling framework and indicate that the model is able to capture a substantial portion of the response variance in voxels across multiple stages of the visual hierarchy.

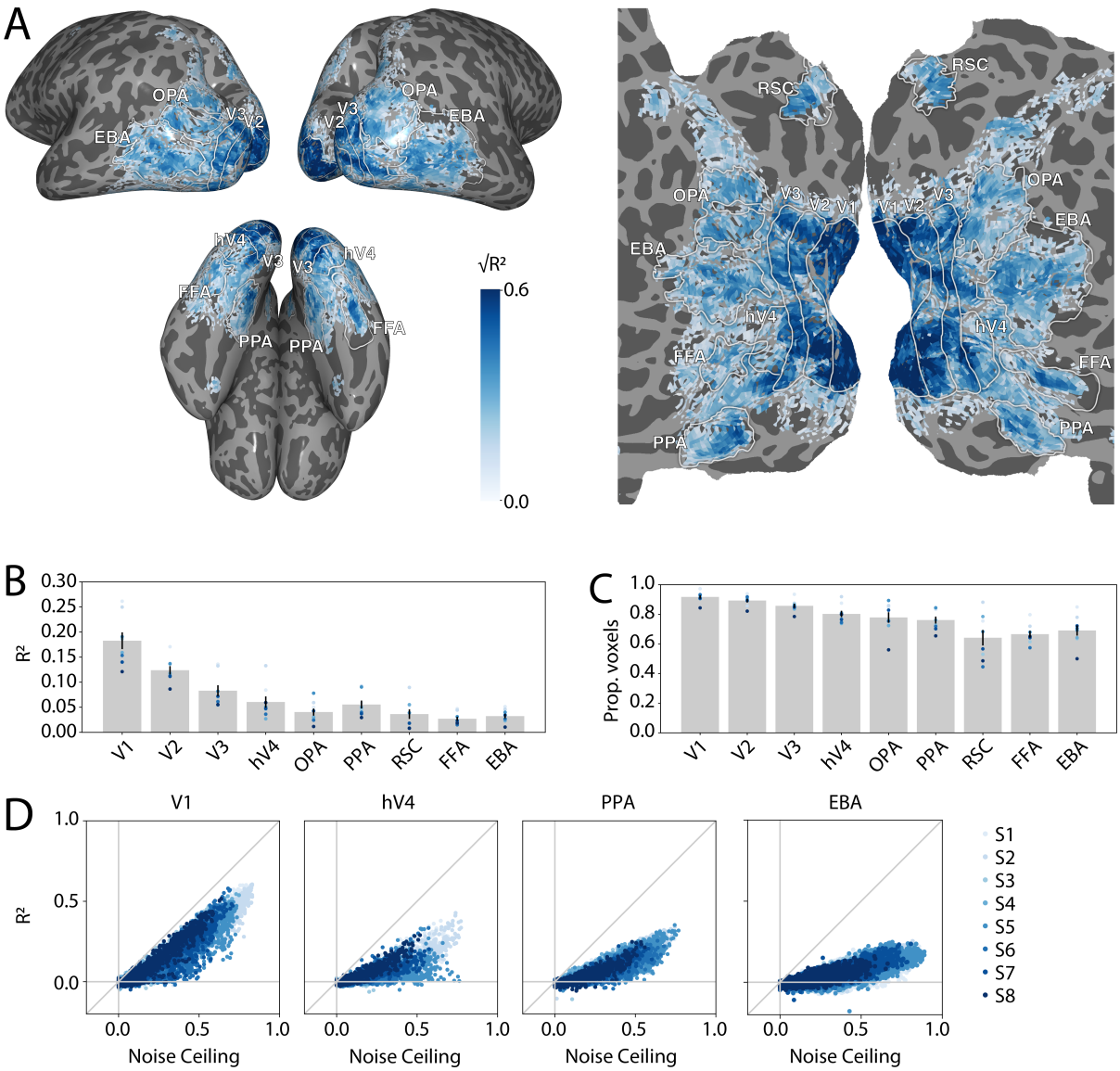

Figure 2: The texture statistics encoding model accurately predicts voxel responses across visual cortex, with highest accuracy in early and mid-level areas. **(A)** Model performance ($\sqrt{R^2}$) is shown on an inflated cortical surface (left) or flattened surface (right) for one example NSD subject, S1 (see Figure 2-1 for the full set of 8 subjects). R^2 was computed using responses to an independent set of images that were not used to fit the model, see *Methods: Model fitting procedure* for details. White outlines and labels indicate the location of functionally-defined retinotopic and category-selective ROIs in each subject; see *Methods: Defining regions of interest (ROIs)* for details. **(B)** Summary of R^2 averaged over voxels within each ROI. **(C)** Proportion of voxels in each ROI where validation set R^2 was significantly higher than chance, evaluated using a permutation test (see *Methods: Permutation testing* for details). In **(B)** and **(C)**, dots in different shades of blue represent individual subjects, bar heights and errorbar represent the mean $\pm$ 1 SEM across subjects. **(D)** R^2 for individual voxels is plotted versus the estimated noise ceiling of the voxels, for four example ROIs. Different shades of blue indicate voxels belonging to different subjects. See *Methods: Noise ceiling estimation* for details on how noise ceiling was estimated.

Given the high overall accuracy of the texture statistics encoding model, we next asked which features were most critical to its predictive performance. To test this, we performed a variance partitioning analysis (see *Methods: Variance partitioning*), where we sub-divided the model features into lower-level and higher-level texture features, and determined both the percentage of variance that was uniquely attributable to each set of

features and the percentage that was shared among the two sets (Figure 3). This analysis revealed key differences among visual areas. First, visualizing the unique variance values on a flattened cortical surface (Figure 3A and Figure 3-1) revealed a gradient from lower to higher visual areas, where voxels in the most posterior portion of occipital cortex tended to have more unique variance explained by the lower-level features (shades of blue), while more anterior voxels, particularly on the lateral surface of the brain, had progressively more variance uniquely explained by the higher-level texture features (shades of red). Averaging the unique variance values across voxels within each ROI further underscored this dissociation between areas: in early retinotopic areas V1-hV4, the lower-level texture features explained more unique variance than did the higher-level texture features, but in higher category-selective visual areas FFA and EBA, this trend reversed, with the higher-level features accounting for more unique variance on average (Figure 3B). Place-selective ROIs OPA, PPA, and RSC showed a more intermediate pattern, with the lower- and higher-level features explaining similar amounts of unique variance. Consistent with these patterns, a bootstrap significance test revealed that the average unique variance explained was significantly higher for the lower-level features than the higher-level features in V1, V2, V3, and hV4, but it was significantly higher for the higher-level features than the lower-level features in EBA (Figure 3-2A; two-tailed p-values computed with a bootstrap test; corrected for multiple comparisons; q=0.01; see *Methods: Variance partitioning*). In addition to significance at the ROI-averaged level, these differences were significant in all individual subjects in V1-V3, in 7 individual subjects in hV4, and in 4 individual subjects in EBA (Table 3-1). Furthermore, more individual voxels in early visual areas had significant unique variance for the lower-level features than for the higher-level features, while more voxels in FFA and EBA had significant unique variance for the higher-level features (Figure 3-2B; one-tailed p-values for single voxels computed using bootstrap test; corrected for multiple comparisons; q=0.01). Another trend evident in this analysis was that across all visual areas, a substantial portion of the variance was shared between the lower-level and higher-level texture features (gray bars in Figure 3B), suggesting some degree of feature overlap between the lower- and higher-level feature spaces.

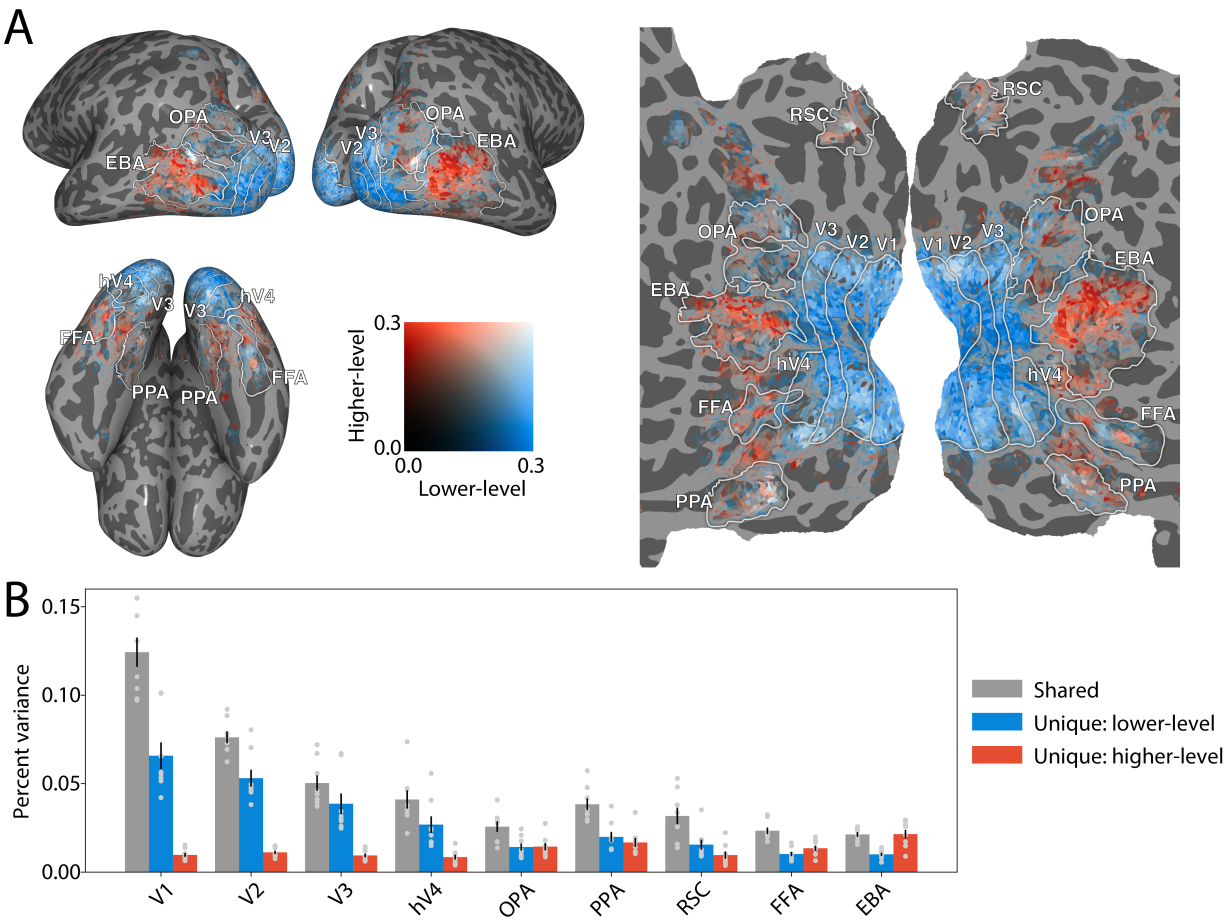

Figure 3: The unique variance explained by the higher-level texture features increases from lower to higher visual areas. Unique variance explained by the lower-level and higher-level features was measured using a variance partitioning analysis, see *Methods: Variance partitioning*. **(A)** Percent variance (units of $\sqrt{R^2}$) uniquely explained by the lower-level (shades of blue) and higher-level (shades of red) features, plotted on an inflated cortical surface from three viewpoints (left) or a flattened cortical surface (right) for one example subject, S1 (see Figure 3-1 for all subjects). Transparent voxels indicate that little variance was uniquely attributable to either the lower- or higher-level features, while white voxels indicate a large proportion of unique variance for both feature types. **(B)** Summary of the variance partitioning results, with R^2 values averaged first over voxels within each ROI and then across subjects. Gray bars indicate variance shared between the lower-level and higher-level features, while blue and red bars indicate variance unique to each set of features. Gray dots indicate individual subjects, bar heights and errorbars indicate mean $\pm$ 1 SEM across subjects. See Figure 3-2 and Table 3-1 for additional analyses examining the significance of the unique variance explained by each feature subset.

The distinction between low- and high-level visual areas in the previous analysis is consistent with the interpretation that these areas represent features at different levels of complexity, with early areas representing more low-level aspects of image structure and higher visual areas representing higher-order statistics. However, another factor that could potentially contribute to this distinction is receptive field size. In our procedure for computing texture statistics features, the pRF size parameter (σ) determines the size of the spatial weighting function used when computing each texture feature (see *Methods: Texture statistics features* for details). This implies that for larger σ values, the higher-order texture statistics features may be more informative, since they incorporate information about a larger portion of the visual field. Since voxels in higher visual cortex tend to have larger σ than voxels in early visual cortex (see Figure 4-1A), it is possible that the ability of higher-level texture features to explain more variance than lower-level features within higher visual areas is driven only by larger receptive field sizes. To evaluate this possibility, we constructed new versions of our encoding models in which σ was fixed at a single value, and then re-fit the entire encoding model with all voxels having the same σ (see *Methods: Fitting with fixed pRF size parameter* for details). Consistent with the difference in estimated receptive field size between areas, this procedure resulted in highest performance in early areas when the fixed σ value was relatively low, but highest performance in higher areas when σ was relatively high (Figure 4-1B). To allow for a balanced comparison across early and higher visual areas, we selected a σ close to the middle of

our range of candidate values, 1.5°, for further inspection.

Importantly, when σ was fixed at 1.5° we replicated the key findings of our original model (Figure 4). We again found that in posterior early visual areas, the lower-level texture features explained a greater proportion of the model variance than the higher-level texture features, but in more anterior areas this pattern began to reverse, with the higher-level texture features explaining a progressively larger percentage of the variance. This pattern was observed across a range of σ values (Figure 4-2), though not for the lowest σ values (note that as described previously, small σ values resulted in low overall R^2 for higher visual areas). The observation that a distinction between the coding properties of low- and high-level visual areas can be recovered even when σ is fixed at a single value for all voxels indicates that the difference between these areas involves a true difference in feature selectivity, not only a difference in the scope of spatial selectivity. This result is also consistent with (Freeman et al., 2013), who found no correlation between the receptive field size of V1 and V2 neurons and their relative sensitivity to higher-level texture statistics.

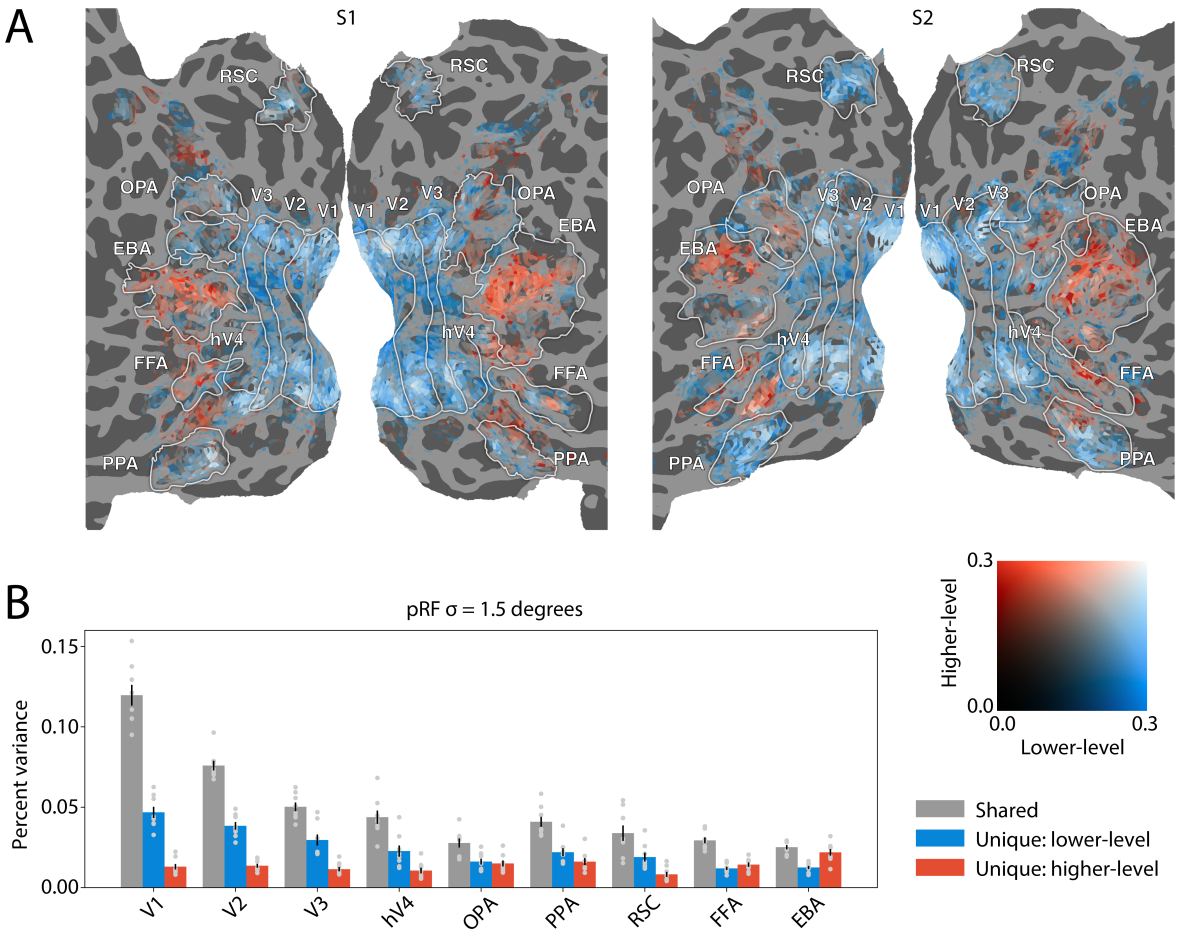

Figure 4: The difference in feature sensitivity between early and higher visual areas is not dependent on differences in receptive field size. We re-fit the entire texture statistics encoding model with a single fixed pRF size (σ=1.5 degrees) for all voxels, see *Methods: Fitting with fixed pRF size parameter* for details. **(A)** Percent variance (units of $\sqrt{R^2}$) unique to the lower-level (shades of blue) and higher-level (shades of red) texture features, shown on a flattened cortical surface for two example subjects. **(B)** The percent of variance that was shared among feature types (gray bars), unique to the lower-level texture features (blue bars), or unique to the higher-level texture features (red bars), averaged over voxels within each ROI. Bar heights and error bars indicate mean ± 1 SEM across subjects, light gray dots indicate individual subjects. See Figure 4-2 for the same analysis for different values of σ.

The previous analyses suggest a dissociation between low and high-level visual areas in terms of which texture statistics features best explain their responses. However, it is not yet clear whether all subsets of the lower- and higher-level texture features contribute equally to model performance in a given brain region, or whether there is redundancy among the feature subsets. Thus, to more precisely determine which texture features were most uniquely predictive of neural activation, we next performed a second variance partitioning analysis where we considered each subset of both the lower-level and higher-level feature groups separately (a total of 10 feature

types; see *Methods: Variance partitioning* and *Methods: Texture statistics features* for details).

As shown in Figure 5 and Figure 6, this analysis revealed that overall model accuracy reflected unique contributions from multiple feature subsets. Across all areas, the feature subset which uniquely explained the largest proportion of the texture model's R^2 was the *energy-mean* features, (Figure 5, top middle panel) which are analogous to a model of complex cell responses in V1. In early areas, a small amount of variance was also explained by the *pixel* features, which include marginal statistics computed from the image luminance histogram (top left panel). A portion of the variance in some early areas was also explained by the *linear-cross-orient* features (bottom left panel), a higher-level feature subset which consists of cross-correlations between the real parts of steerable pyramid bands with different orientations. These features may capture higher-order structure generated from the combination of different orientations, such as contour junctions. Moving into more anterior regions, other higher-level features began to explain a larger proportion of the variance. The *linear-auto* features, which are computed by taking the spatial autocorrelation of the steerable pyramid's lowpass reconstructions at each scale and the highpass residual band, explained a proportion of variance across several high-level visual areas, including voxels in and near the category-selective areas (center panel). The spatial autocorrelations in this feature subset may allow it to capture spatially repeating textural elements in the images. Other feature subsets which explained a notable amount of unique variance in higher-level visual cortex included those computed from higher-order statistics of the steerable pyramid magnitude features, including the *energy-cross-scale* features, as well as the *energy-cross-orient* features. The spatial pattern of unique variance explained by each feature type also suggested some heterogeneity within individual ROIs for this example subject. For example, the *energy-cross-orient* features explained a large amount of unique variance in only the superior portion of EBA (center right panel), and the *linear-cross-orient* features explained a sizeable amount of variance in only the posterior portion of PPA (bottom left).

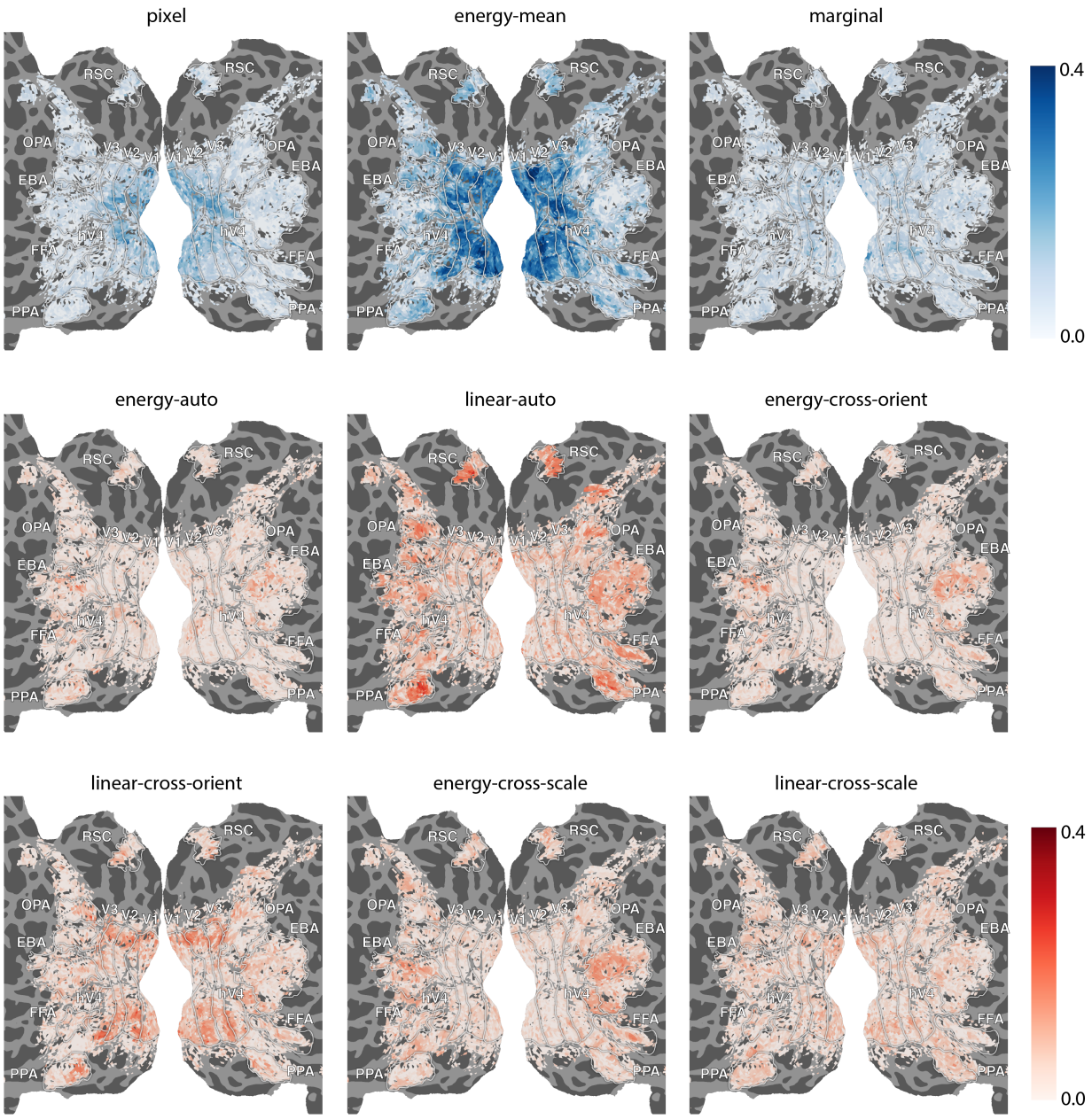

Figure 5: The spatial distribution of neural populations sensitive to each subset of texture model features, shown for one representative NSD subject (S1). Each panel shows the percentage of encoding model variance (units of $\sqrt{R^2}$) uniquely explained by each individual feature type, shown on a flattened cortical surface (see *Methods: Texture statistics features* for details on feature types). The maps in shades of blue (top row) correspond to subsets of the lower-level texture features, while maps in shades of red (bottom two rows) correspond to subsets of higher-level texture features. Note that the *linear-mean* features are not plotted here, as they explained little unique variance in any ROI (see Figure 6).

To quantify these effects, we averaged the unique variance across voxels in each ROI (Figure 6, Figure 6-1, Table 6-1). Again, this analysis demonstrated a primary role for the *energy-mean* features in all ROIs, with this feature subset dominating the most strongly in early visual areas. The unique variance explained by the *energy-mean* features was significant at the subject-averaged level in all ROIs except EBA, and was significant in all individual subjects in V1, V2, V3, and hV4 (one-tailed p-values computed using bootstrap test; corrected for multiple comparisons; q=0.01; see *Methods: Variance partitioning* for details). Though the absolute amount of unique variance contributed by other feature subsets was modest in comparison to the *energy-mean* features, other feature subsets nonetheless contributed a significant amount of unique variance in each ROI. The *pixel* features made significant contributions in several areas, with highest unique variance contributed in V1 and V2 (all 8 subjects individually significant in V1 and V2). As suggested in the previous paragraph, the *linear-cross-orient* features also contributed to the variance explained in many areas, with unique variance explained significant at the subject-averaged level in all ROIs (significant in all 8 subjects in V1 and V2). Another higher-

level feature subset that stood out was the *linear-auto* features, which explained a significant amount of unique variance in V3 as well as OPA, PPA, RSC, FFA and EBA. These results suggest that the performance of the texture statistics encoding model reflects contributions from multiple feature subsets within the model, with the relative importance of these feature subsets varying among visual areas. This indicates that areas across the visual cortex may encode complementary aspects of image texture, resulting in a population representation that spans a large portion of the full texture statistics feature space.

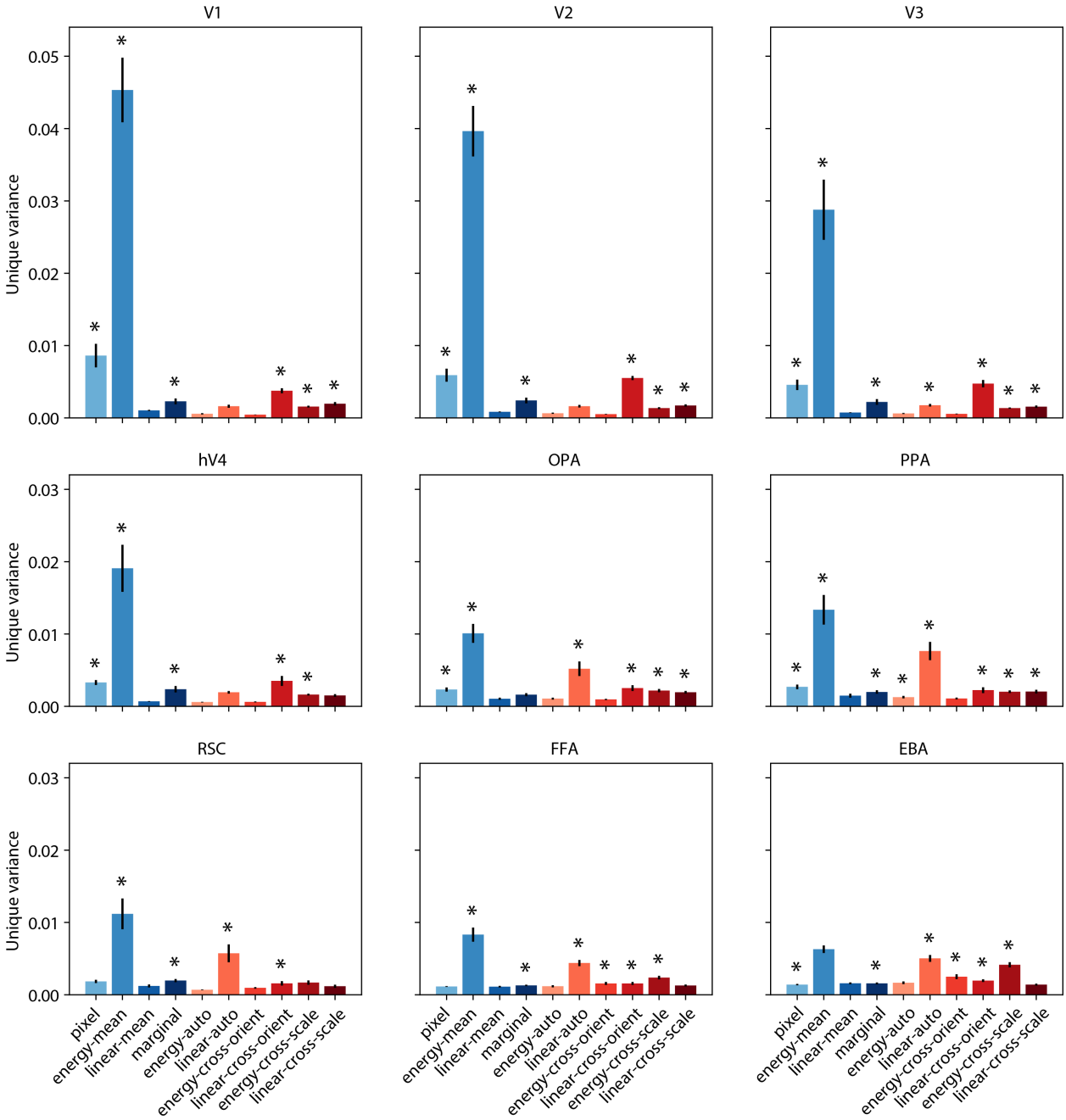

Figure 6: Results of variance partitioning analysis across all feature types, summarized at the ROI level. Each panel shows the proportion of model variance uniquely explained by each feature type, averaged over voxels within a single ROI for all subjects. Bar heights and errorbars indicate the mean ± 1 SEM across subjects, light gray dots indicate means for individual subjects. Bars in shades of blue indicate feature subsets belonging to the "lower-level" group, while bars in shades of red indicate feature subsets belonging to the "higher-level" group. Asterisks (*) above each bar indicate that the unique variance explained by a feature subset (averaged across all subjects) was significantly greater than zero (one-tailed p-values computed using a bootstrap test; corrected for multiple comparisons; q=0.01; see *Methods: Variance partitioning* for details). See Figure 6-1 for a plot of the confidence intervals generated during the bootstrap significance test, and see Table 6-1 for the number of individual subjects in which unique variance was significant for each feature subset.

Discussion

Intermediate-level feature representations are an essential component of the human visual hierarchy, providing a link between low-level image properties and high-level conceptual information. However, our understanding of the features and computations that contribute to mid-level representations, and how these representations may differ across levels of the visual hierarchy, has remained incomplete. Here, we addressed this knowledge gap by constructing a voxelwise encoding model based on a set of image-computable texture statistics features ("P-S statistics"; Portilla and Simoncelli, 2000). We demonstrated, in early retinotopic areas as well as in higher category-selective regions, that our model could generate accurate predictions of voxel responses to individual natural scene images that were not used during model fitting. We further showed that the features contributing most to the predictive accuracy of the model differed depending on position in the visual hierarchy: lower-level texture features explained more unique variance in early visual cortex, while higher-level texture features progressively explained a larger amount of unique variance in more anterior regions. These results illuminate the nature of the mid-level representations instantiated in the human visual system, and they further highlight that mid-level representations may play an essential role in processing even at late stages of the ventral visual stream.

Past work has measured the sensitivity to P-S statistics in V1, V2, and V4 in the primate brain, finding results that are broadly consistent with ours. First, multiple studies have shown that spectral features (i.e., the *energy-mean* features in our model), tend to have the largest regression model weights of any P-S feature type in V1, V2, and V4 (Hatanaka et al., 2022; Okazawa et al., 2017, 2015); this is consistent with the results of our variance partitioning analyses (Figure 6 and Figure 6-1), which showed the highest unique variance explained for the *energy-mean* features in every early visual ROI. Beyond the *energy-mean* features, we found that the *pixel* features had the second-largest amount of unique variance explained in V1, which is consistent with some past work (Hatanaka et al., 2022). Our finding that higher-level feature sensitivity increased from early to high-level areas has also been suggested previously. Past studies found that the contributions of higher-order texture statistics are larger on average for neurons in V4 than V2 (Okazawa et al., 2017), and larger for V4 than V1 (Hatanaka et al., 2022). These results are consistent with our finding that the percent variance uniquely explained by the higher-level texture features was higher, relative to the total variance explained by the model, in hV4 relative to V1 or V2 (Figure 3B). Importantly, our results expand upon these findings by demonstrating the continuous increase in feature complexity not only in a small number of early visual regions but across a wide range of areas in the visual system.

One aspect of our results that has more mixed agreement with past work is the relative importance of the different higher-level texture feature subsets. In early areas V1-hV4, we found that the higher-level texture feature subset which explained the most unique variance on average was the *linear-cross-orient* features. One past study also found relatively high average weights for this set of features (Hatanaka et al., 2022), along with the *linear-cross-scale* features. However other studies did not include this feature subset in their models (Okazawa et al., 2017, 2015), and their results instead suggested a key role for the *energy-cross-position* and *energy-cross-scale* features. These discrepancies may be the result of modeling differences across studies, such as the inclusion of certain feature subsets, particularly if some subsets carry redundant information with one another. Differences in stimulus type may also be a contributing factor to these discrepancies as two of these past studies (Okazawa et al., 2017, 2015) used synthetic texture stimuli as opposed to natural scene images, and sensitivity to higher-order texture features may be affected by the presence of additional structure included in natural scene images. Another potential contributing factor is a difference in the measurement precision between fMRI and single-unit recordings. For example, if groups of neurons having different types of selectivity for higher-order texture features are spatially intermingled in visual cortex, some aspects of their selectivity may not be detectable at the voxel level. On the other hand, the fact that a significant amount of unique variance was explained by the higher-level texture features in many voxels (Figure 3-2B) argues against the idea that higher-level statistics were inaccessible at the voxel level.

Although the lower-level texture features, in particular the *energy-mean* features, explained significantly more unique variance in V1 than the higher-level texture features (Figure 3-2A), a small but significant amount of unique variance in V1 was explained by the higher-level texture features (significant in all 8 individual subjects; Table 3-1). In particular, V1 (as well as V2, V3, and hV4) had a significant amount of unique variance explained by the *linear-cross-orient* features, which include information about the relationships between different orientations in the image. This result is somewhat surprising given that the responses of both V1 simple cells and V1 complex cells should theoretically be captured by the lower-level features of the model (i.e., the *linear-mean* and *energy-mean* features), and that past work has suggested that V1 neurons are not sensitive to the higher-order correlation statistics of the P-S model (Freeman et al., 2013). Our results may indicate that, consistent with some other work, V1 neurons are sensitive to information that cannot be captured by simple, spatially

localized spectral features (Carandini et al., 2005). For example, one study showed that V1 neurons presented with simple texture stimuli can signal the presence of some higher-order statistical correlations (Purpura et al., 1994). Other work has suggested that V1 neurons may be sensitive to contextual influences reflecting modulation from outside the classical RF region, leading to response properties such as figure-ground separation, sensitivity to illusory contours, or sensitivity for perceived size (Albright & Stoner, 2002; Murray et al., 2006). Such contextual sensitivity could be one mechanism for the appearance of selectivity for higher-order image structure in our results, in particular the sensitivity to cross-orientation information indicated by the *linear-cross-orient* features. Given the limited temporal resolution of fMRI, we cannot distinguish whether the sensitivity to higher-level texture statistics in V1 is due to feedforward processing, lateral interactions within V1, or feedback from higher visual regions.

Beyond early visual cortex, we found evidence for selective sensitivity to higher-level texture features in more anterior areas of the visual system, including voxels in and around category-selective visual regions such as EBA (Figure 3). This result is consistent with a previous study demonstrating that the fusiform gyrus and the middle occipital gyrus exhibit larger fMRI activation for simple texture stimuli including higher-order correlations than those lacking these correlations (Beason-Held et al., 1998). Our results also revealed that, when features were broken down on a more fine-grained level, the *linear-auto* texture features were among the most predictive features for higher visual areas, explaining a small but significant amount of unique variance at the ROI-averaged level in OPA, PPA, RSC, FFA, and EBA (Figure 6 and Figure 6-1). This feature subset likely captures image periodicity at various scales, which is consistent with past work suggesting ventral visual areas VO1 and LOC are sensitive to the degree of spatial regularity in synthetic texture stimuli (Kohler et al., 2016). Other higher-level feature subsets also contributed modest but significant amounts of unique variance in higher visual areas, with the *energy-cross-scale* features having the highest average unique variance in EBA. The cross-scale correlations included in computing these features may allow them to capture information about edges and lines in images. Additionally, although they explained more unique variance in early areas than higher visual areas, the *linear-cross-orient* features also explained a significant amount of unique variance in all ROIs examined, suggesting they may play a key role in mid-level computations throughout the visual system. In particular, these features may encode higher-order conjunctions between orientations such as contour junctions, angles, and curvature, and thus may play an important role in the computation of higher-level object and scene categories (Walther & Shen, 2014).

These patterns should be considered in light of the fact that the overall performance of the model in more anterior regions was relatively poor compared to early areas, with R^2 only reaching a fraction of the noise ceiling for most voxels (Figure 2D). This indicates that the texture statistics model does not fully capture the range of image features to which voxels in higher visual cortex are sensitive. This conclusion is consistent with a study which compared V4 and inferotemporal (IT) cortex in macaques (Rust & DiCarlo, 2010), and found that intact images were more discriminable in IT than scrambled images with matched P-S statistics, while in V4 the difference between these stimulus classes was smaller. This implies that IT is sensitive to certain high-level aspects of natural image structure that are not captured by the P-S statistics. Similarly, an fMRI study using scrambled images that matched the P-S statistics of object images showed that the category selectivity in higher visual cortex induced by scrambled images is weaker than that induced by intact images (Long et al., 2018), again indicating sensitivity to high-level features that go beyond texture statistics. Thus, perhaps unsurprisingly, the texture statistics encoding model does not appear to provide a complete account of the response properties of higher visual cortex. However, our results do support the idea that intermediate visual features are reliably encoded within high-level visual areas, and that modeling these mid-level features can contribute to a more detailed understanding of the representations in these regions.

Developing an accurate model of mid-level computations in the visual system is a key step toward understanding how complex representations are generated from low-level feature maps. Here, we demonstrated that a representational space based on texture statistics features can be used to predict voxel responses to arbitrary natural images, and that multiple subsets of features in the model provide unique contributions to its predictive accuracy. Although model performance was highest in early and intermediate visual areas, our results suggest that texture statistics features are represented in category-selective visual cortex as well. This finding paves the way for future work to examine how such representations may contribute to higher cognitive processes such as the recognition of complex objects and scenes.

Acknowledgements

This research was funded by a Distinguished Postdoctoral Fellowship from the Carnegie Mellon Neuroscience Institute to MMH. Collection of the NSD dataset was supported by NSF IIS-1822683 and NSF IIS-1822929.

References

Albright, T. D., & Stoner, G. R. (2002). Contextual influences on visual processing. *Annual Review of Neuroscience*, *25*, 339–379.

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2021). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*, 116–126.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*, 13–13.

Balas, B. J. (2006). Texture synthesis and perception: Using computational models to study texture representations in the human visual system. *Vision Research*, *46*, 299–309.

Baumgartner, E., & Gegenfurtner, K. R. (2016). Image statistics and the representation of material properties in the visual cortex. *Frontiers in Psychology*, *7*, 1185.

Beason-Held, L. L., Purpura, K. P., Krasuski, J. S., Maisog, J. M., Daly, E. M., Mangot, D. J., Desmond, R. E., Optican, L. M., Schapiro, M. B., & Vanmeter, J. W. (1998). Cortical regions involved in visual texture perception: A fmri study. *Cognitive Brain Research*, *7*, 111–118.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*, 289–300.

Benson, N. C., Jamison, K. W., Arcaro, M. J., Vu, A. T., Glasser, M. F., Coalson, T. S., Essen, D. C. V., Yacoub, E., Ugurbil, K., Winawer, J., & Kay, K. (2018). The human connectome project 7 tesla retinotopy dataset: Description and population receptive field analysis. *Journal of Vision*, *18*, 23–23.

Bergen, J. R., & Landy, M. S. Computational modeling of visual texture segregation (M. S. Landy & J. A. Movshon, Eds.). In: In *Computational models of visual processing* (M. S. Landy & J. A. Movshon, Eds.). Ed. by Landy, M. S., & Movshon, J. A. The MIT Press, 1991, pp. 253–271.

Bracci, S., Ritchie, J., & de Beeck, H. O. (2017). On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia*, *105*, 153–164.

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L., & Rust, N. C. (2005). Do we know what the early visual system does? *The Journal of Neuroscience*.

Connor, C. E., Brincat, S. L., & Pasupathy, A. (2007). Transformation of shape information in the ventral pathway. *Current Opinion in Neurobiology*, *17*, 140–147.

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, *4*, 2051–2062.

Dumoulin, S. O., & Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *NeuroImage*, *39*, 647–660.

Duncan, R. O., & Boynton, G. M. (2003). Cortical magnification within human primary visual cortex correlates with acuity thresholds. *Neuron*, *38*, 659–671.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*, 1195–1204.

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

Gao, J. S., Huth, A. G., Lescroart, M. D., & Gallant, J. L. (2015). Pycortex: An interactive surface visualizer for fmri. *Frontiers in Neuroinformatics*, *9*.

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, *15*, 536–548.

Groen, I. I. A., Silson, E. H., & Baker, C. I. (2017). Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*.

Groen, I. I., Ghebreab, S., Lamme, V. A., & Scholte, H. S. (2012). Spatially pooled contrast responses predict neural and perceptual similarity of naturalistic image categories. *PLOS Computational Biology*, *8*, e1002726.

Güçlü, U., & van Gerven, M. A. (2014). Unsupervised feature learning improves prediction of human brain activity in response to natural images. *PLOS Computational Biology*, *10*, e1003724.

Hatanaka, G., Inagaki, M., Takeuchi, R. F., Nishimoto, S., Ikezoe, K., & Fujita, I. (2022). Processing of visual statistics of naturalistic videos in macaque visual areas v1 and v4. *Brain Structure and Function*, *227*, 1385–1403.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*, 106.

Huth, A. G., Heer, W. A. D., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*, 453–458.

Kohler, P. J., Clarke, A., Yakovleva, A., Liu, Y., & Norcia, A. M. (2016). Representation of maximally regular textures in human visual cortex. *Journal of Neuroscience*, *36*, 714–729.

Lescroart, M. D., & Gallant, J. L. (2019). Human scene-selective areas represent 3d configurations of surfaces. *Neuron*, *101*, 178–192.e7.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8693 LNCS*, 740–755.

Long, B., Konkle, T., Cohen, M. A., & Alvarez, G. A. (2016). Mid-level perceptual features distinguish objects of different real-world sizes. *Journal of Experimental Psychology: General*, *145*, 95–109.

Long, B., Störmer, V. S., & Alvarez, G. A. (2017). Mid-level perceptual features contain early cues to animacy. *Journal of Vision*, *17*, 20.

Long, B., Yu, C. P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, E9015–E9024.

Murray, S. O., Boyaci, H., & Kersten, D. (2006). The representation of perceived angular size in human primary visual cortex. *Nat Neurosci*, *9*(3), 429–434.

Naselaris, T., & Kay, K. N. (2015). Resolving ambiguities of mvpa using explicit models of representation. *Trends in Cognitive Sciences*, *19*, 551–554.

Nasr, S., Echavarria, C. E., & Tootell, R. B. (2014). Thinking outside the box: Rectilinear shapes selectively activate scene-selective cortex. *Journal of Neuroscience*, *34*, 6721–6735.

Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). Voxelwise encoding models with non-spherical multivariate normal priors. *NeuroImage*, *197*, 482–492.

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.

Okazawa, G., Tajima, S., & Komatsu, H. (2015). Image statistics underlying natural texture selectivity of neurons in macaque v4. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, E351–E360.

Peirce, J. W. (2015). Understanding mid-level representations in visual processing. *Journal of Vision*, *15*, 5–5.

Portilla, J., & Simoncelli, E. P. (2000). Parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*, 49–71.

Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., & Kay, K. N. (2022). Glmsingle: A toolbox for improving single-trial fmri response estimates. *bioRxiv*, 2022.01.31.478431.

Purpura, K. P., Victor, J. D., & Katz, E. (1994). Striate cortex extracts higher-order spatial correlations from visual textures. *Proceedings of the National Academy of Sciences of the United States of America*, *91*, 8482–8486.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision*, *12*, 14–14.

Rust, N. C., & DiCarlo, J. J. (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area v4 to it. *Journal of Neuroscience*, *30*, 12978–12995.

Serences, J. T., & Saproo, S. (2012). Computational advances towards linking bold and behavior. *Neuropsychologia*, *50*, 435–446.

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

St-Yves, G., & Naselaris, T. (2017). The feature-weighted receptive field: An interpretable encoding model for complex feature spaces. *NeuroImage*, *180*, 188–202.

Stigliani, A., Weiner, K. S., & Grill-Spector, K. (2015). Temporal processing capacity in high-level visual cortex is domain specific. *Journal of Neuroscience*, *35*, 12412–12424.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, *5*, 682–687.

Walther, D. B., & Shen, D. (2014). Non-accidental properties underlie human categorization of complex natural scenes. *Psychological science*, *25*, 851.

Wang, L., Mruczek, R. E., Arcaro, M. J., & Kastner, S. (2015). Probabilistic maps of visual topography in human cortex. *Cerebral Cortex*, *25*, 3911–3931.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLOS ONE*, *9*, e112575.

Extended Data

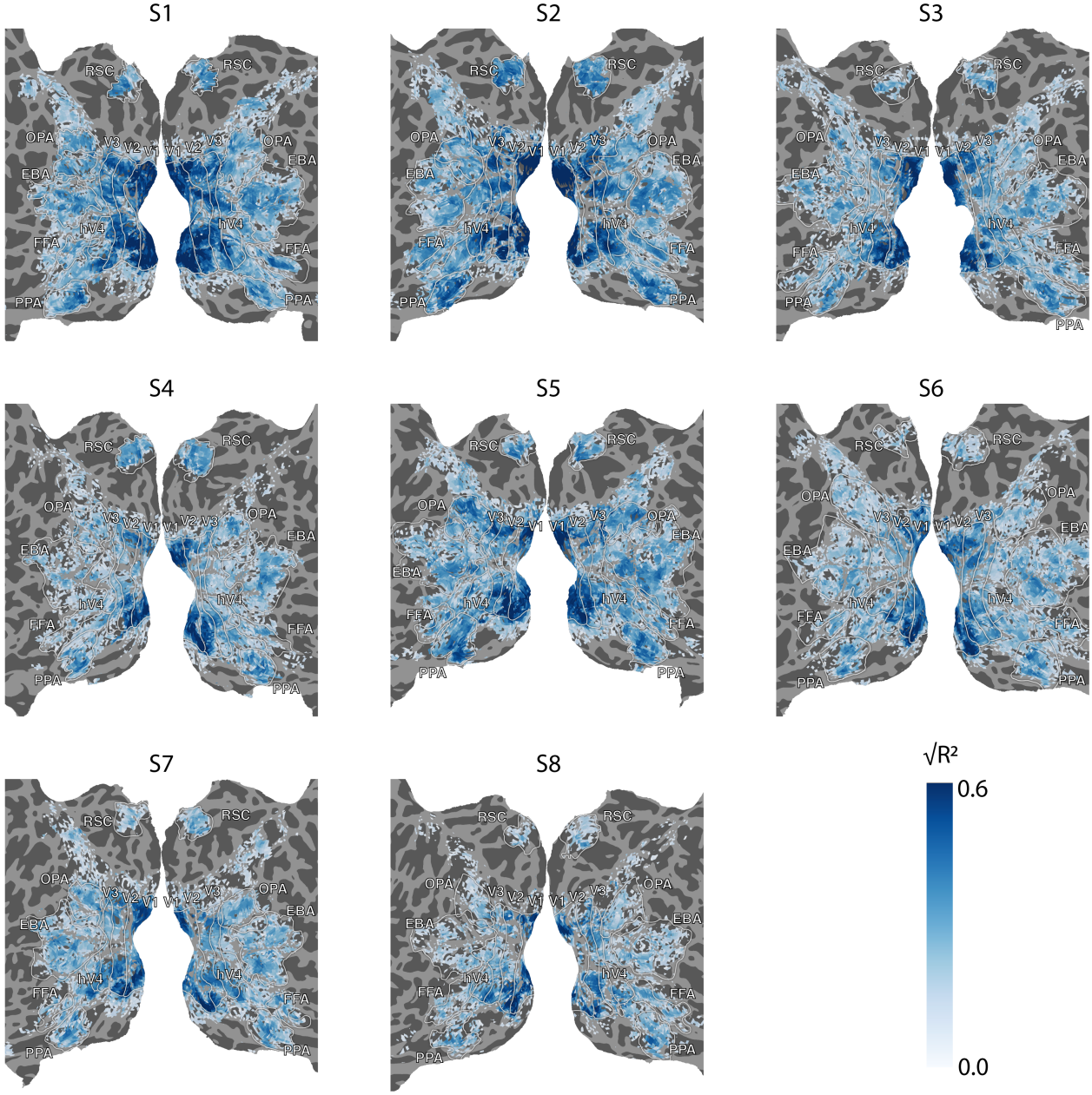

Figure 2-1: Validation set accuracy of the texture statistics encoding model ($\sqrt{R^2}$) plotted on a flattened cortical surface, for all 8 NSD subjects. White outlines and labels indicate the location of functionally-defined retinotopic and category-selective ROIs in each subject; see *Methods: Defining regions of interest (ROIs)* for details.

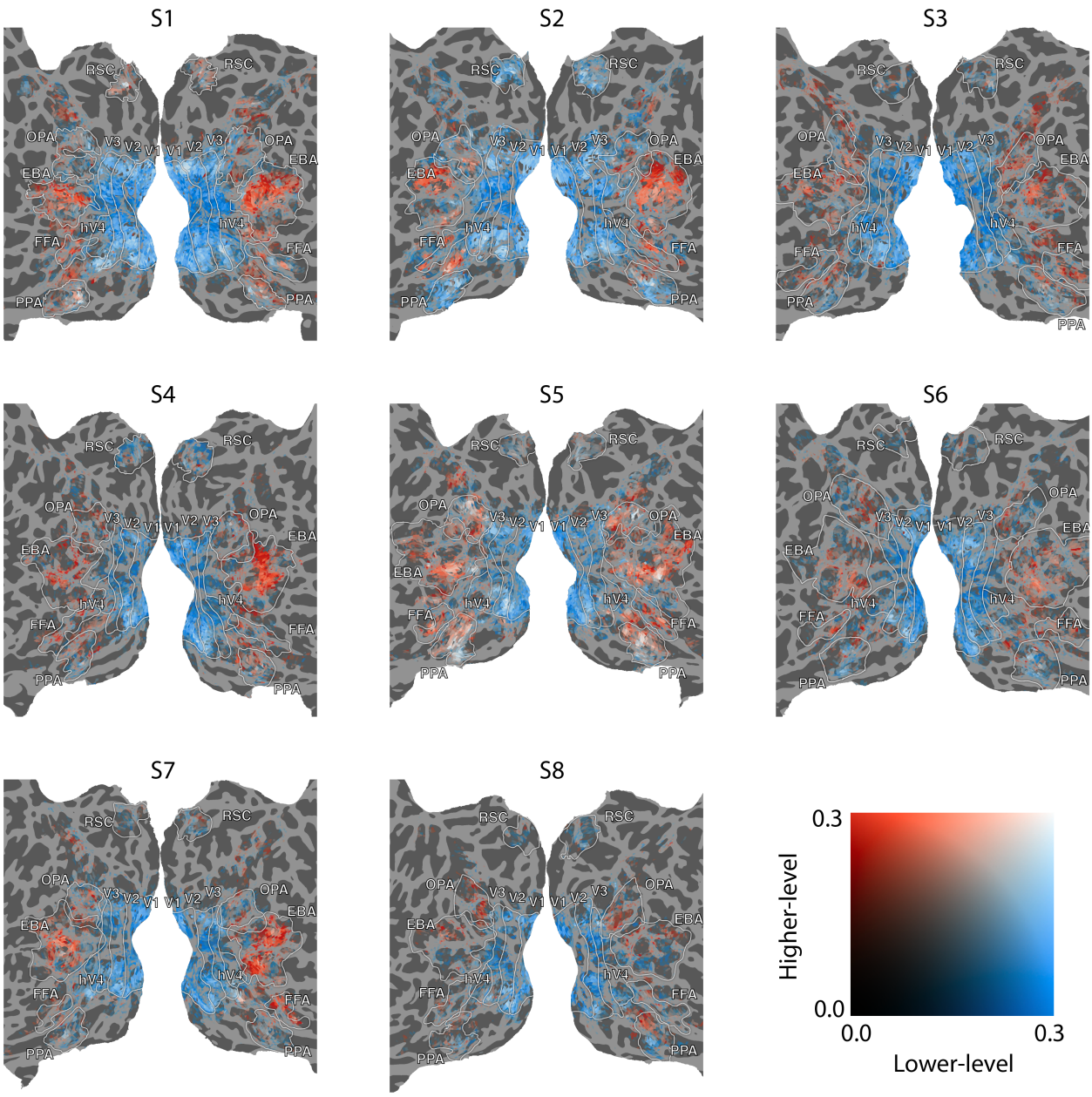

Figure 3-1: Comparison of unique variance explained by the lower-level and higher-level texture features, for all 8 subjects. The variance uniquely explained by lower-level and higher-level texture features was estimated using a variance partitioning analysis (see *Methods: Variance partitioning* for details). Values are in units of $\sqrt{R^2}$, and are plotted using a 2D coloring scheme, where shades of blue and red indicate the percent of variance attributable to the lower-level and higher-level features, respectively. Transparent voxels indicate that little variance was uniquely attributable to either the lower- or higher-level features, while white voxels indicate a large proportion of unique variance for both feature types.

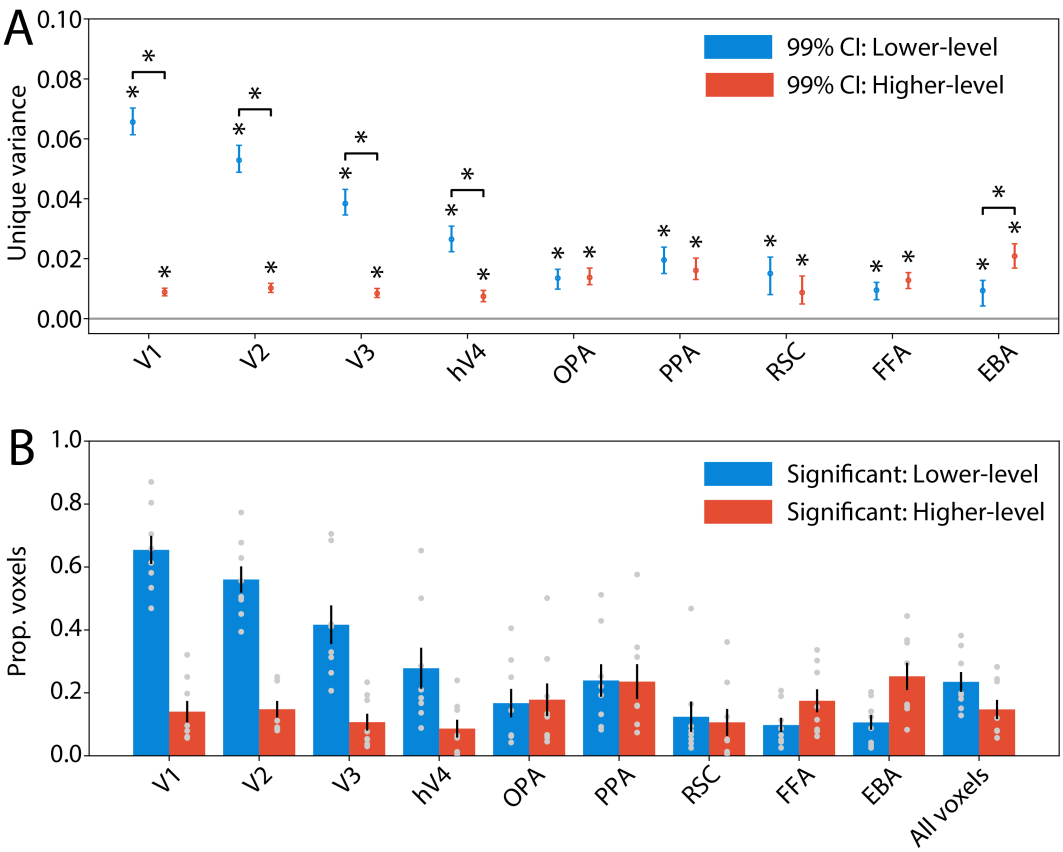

Figure 3-2: Additional analyses examining the significance of unique variance explained by the lower- and higher-level texture features. **(A)** Confidence intervals (99%) for the average unique variance explained by each set of texture model features, obtained by bootstrapping the images when computing R^2 (see *Methods: Variance partitioning* for details). Asterisks (*) above intervals indicate that the average unique variance was significantly greater than zero (one-tailed p-values computed using a bootstrap test; corrected for multiple comparisons; q=0.01), asterisks above pairs of bars indicate a significance difference between the lower-level and higher-level features (two-tailed p-values computed using a bootstrap test; corrected for multiple comparisons; q=0.01). **(B)** Proportion of individual voxels in each ROI that had a significant amount of variance uniquely explained by the lower-level (blue bars) and higher-level (red bars) texture features (one-tailed p-values computed using a bootstrap test; corrected for multiple comparisons; q=0.01; see *Methods: Variance partitioning* for details). Gray dots indicate individual subjects, bar heights and error bars indicate mean ±1 SEM across subjects. The rightmost bars ("All voxels") indicate the proportion of voxels that were significant across the entire analyzed visual cortex region, without regard to ROI definitions.

	V1	V2	V3	hV4	OPA	PPA	RSC	FFA	EBA
lower-level	8	8	8	8	6	8	2	5	4
higher-level	8	8	7	5	8	8	3	7	8
difference	8	8	8	7	0	1	0	0	4

Table 3-1: First two rows show the number of individual subjects (out of 8 total) in which the unique variance explained by the lower-level and higher-level texture features was significantly greater than zero, for each ROI (one-tailed p-values computed using a bootstrap test; corrected for multiple comparisons; q=0.01; see *Methods: Variance partitioning* for details). Last row shows the number of subjects for which the difference between the unique variance explained by the lower- and higher-level texture features was significantly different from zero (two-tailed p-values computed using a bootstrap test; corrected for multiple comparisons; q=0.01; see *Methods: Variance partitioning* for details). Where significant differences were detected, their signs were as follows: lower-level > higher-level in V1, V2, V3, hV4, PPA; higher-level > lower-level in EBA.

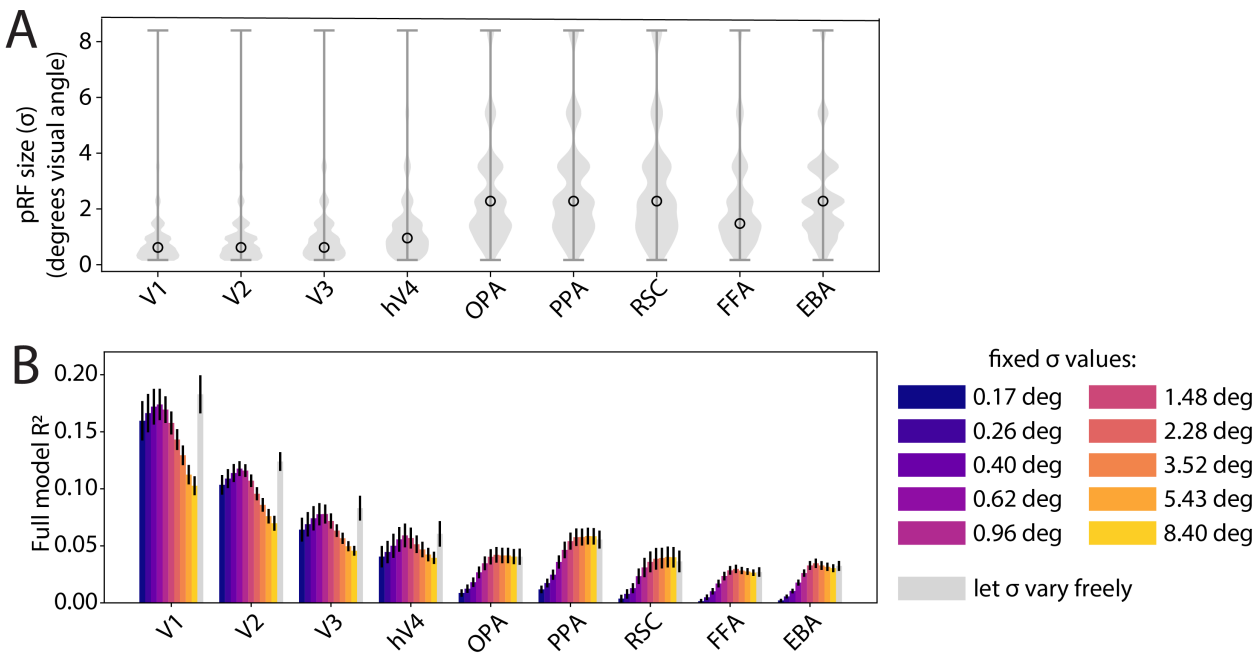

Figure 4-1: Differences in pRF size (σ) across ROIs, and effect of restricting σ on model accuracy. **(A)** Distribution of pRF σ across voxels in each ROI, computed using the model in which σ was allowed to vary across voxels. Distributions include voxels pooled across all subjects, open circle indicates the median. **(B)** Overall accuracy (R^2) of models in which pRF size (σ) was fixed at a single value for all voxels (see *Methods: Fitting with fixed pRF size parameter*), averaged across voxels in each ROI. Each colored bar represents a different σ value; gray bar represents the model in which sigma was allowed to vary across voxels (similar to Figure 2B). Single-voxel R^2 was averaged over voxels within each ROI, and then across subjects. Bar height and errorbars indicate the mean ± SEM across subjects.

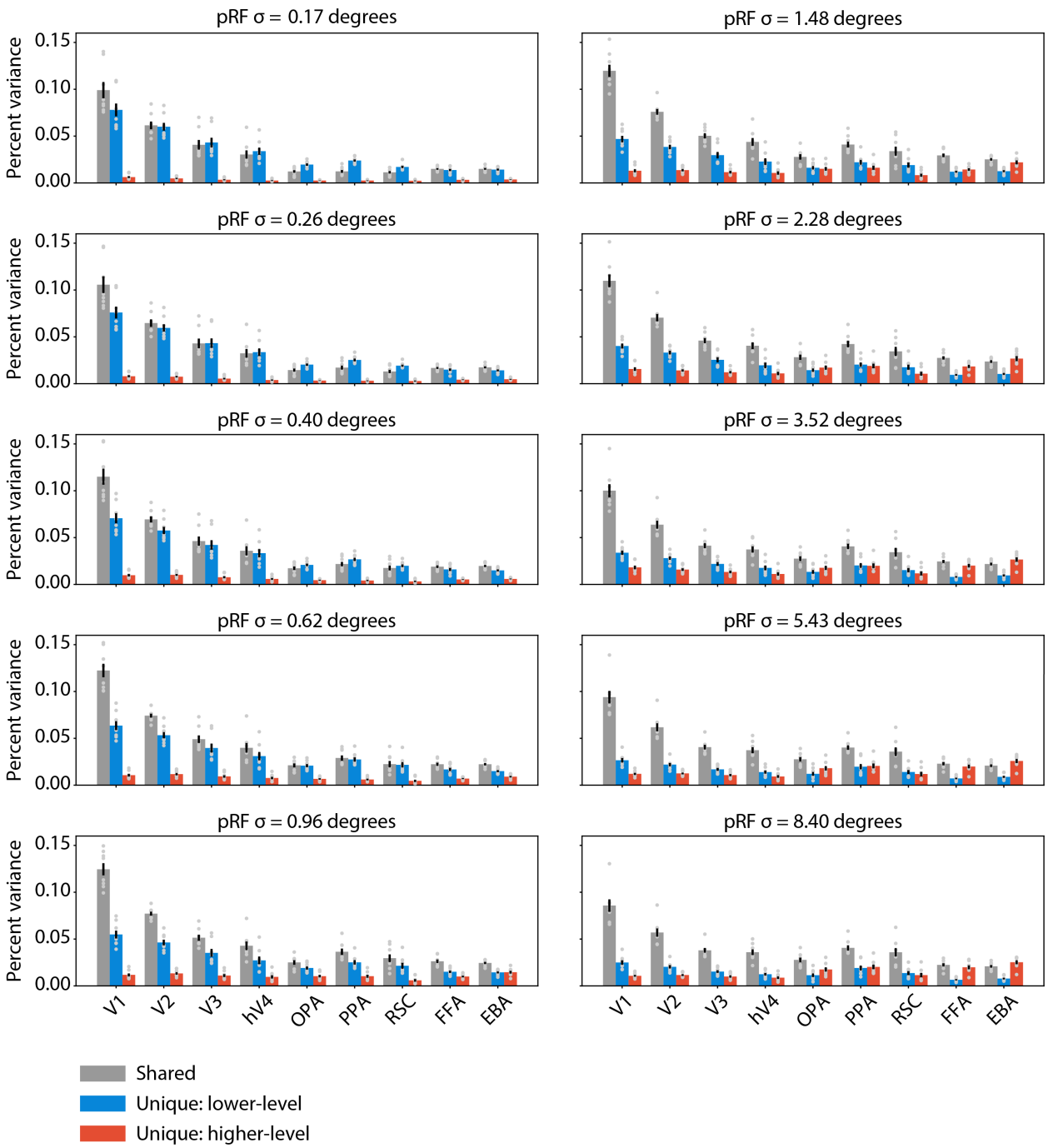

Figure 4-2: Variance partitioning analysis for lower-level and higher-level texture features, evaluated using models where pRF size (σ) was fixed at a single value for all voxels (see *Methods: Fitting with fixed pRF size parameter* and *Methods: Variance partitioning* for details). In each plot, gray, blue, and red bars indicate the proportion of variance that is shared, unique to the lower-level texture features, and unique to the higher-level texture features, respectively. Values are averaged across voxels within each ROI, and then across subjects. Bar height and errorbars indicate the mean ± SEM across subjects, light gray dots indicate individual subjects. Each individual panel represents the same analysis performed for a different σ value. Note that smaller σ resulted in poor performance in higher visual cortex, see Figure 4-1B.

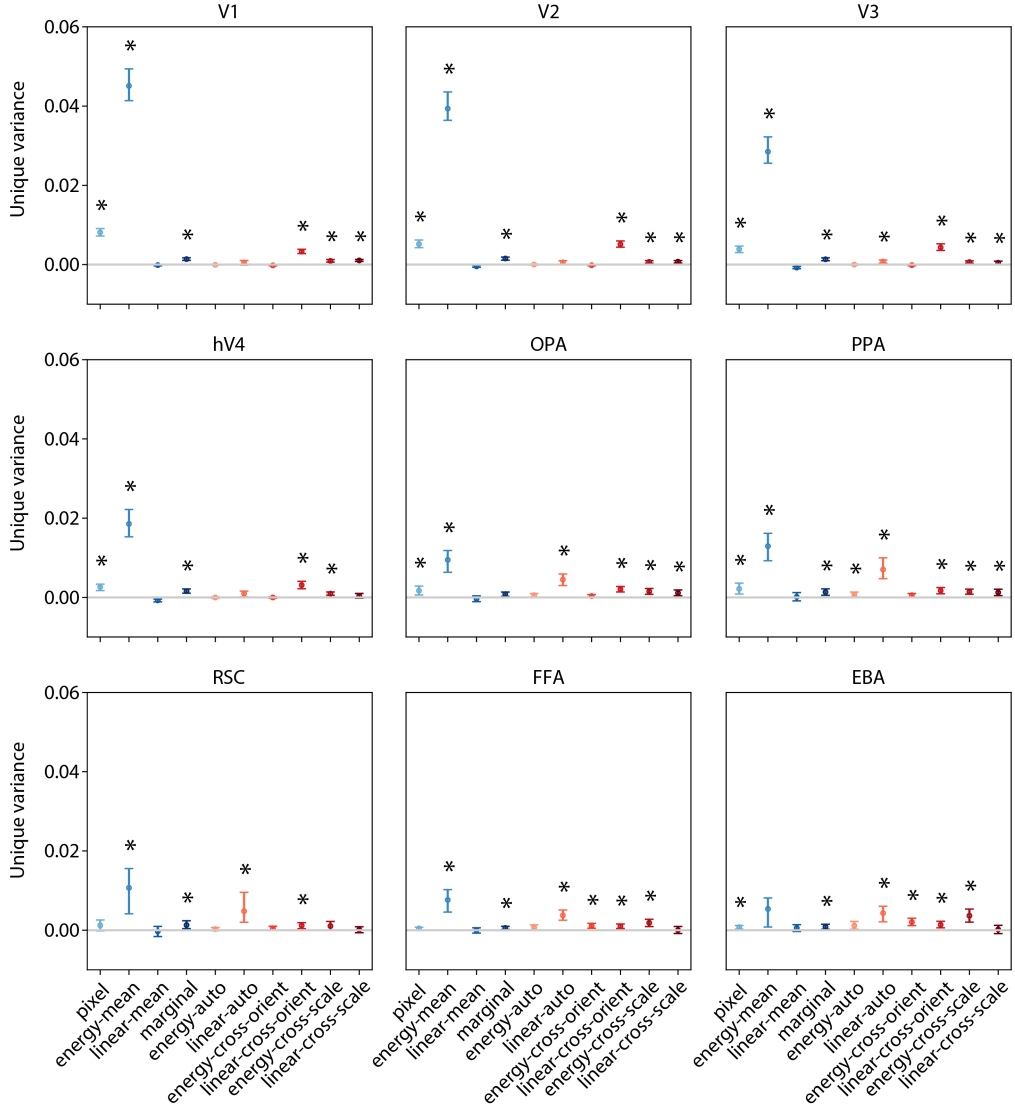

Figure 6-1: Confidence intervals (99%) for the average unique variance explained by each feature subset in the texture model, obtained by bootstrapping over images when computing R^2 (see *Methods: Variance partitioning* for details). Shades of blue indicate features included in the "lower-level" group, while shades of red indicate features included in the "higher-level" group. Asterisks (*) above each bar indicate that the unique variance explained was significantly greater than zero (one-tailed p-values computed using bootstrap test; p-values corrected for multiple comparisons; q=0.01; see *Methods: Variance partitioning* for details).

	V1	V2	V3	hV4	OPA	PPA	RSC	FFA	EBA
pixel	8	8	7	4	2	1	0	0	2
energy-mean	8	8	8	8	3	7	2	4	1
linear-mean	0	0	0	0	0	0	0	0	0
marginal	3	5	3	3	2	2	0	0	0
energy-auto	1	1	0	0	0	0	0	0	0
linear-auto	1	0	1	0	6	5	2	4	4
energy-cross-orient	0	0	0	0	0	0	0	2	2
linear-cross-orient	8	8	7	6	4	3	0	2	1
energy-cross-scale	2	0	1	3	2	1	0	0	4
linear-cross-scale	4	2	2	1	0	1	0	0	0

Table 6-1: Number of individual subjects (out of 8 total) in which the unique variance explained by each subset of texture features was significantly greater than zero, for each ROI (one-tailed p-values computed using a bootstrap test; corrected for multiple comparisons; q=0.01; see *Methods: Variance partitioning* for details).