RESEARCH ARTICLE

Exploring spatiotemporal neural dynamics of the human visual cortex

Ying Yang¹ | Michael J. Tarr^{1,2} | Robert E. Kass^{1,3,4} | Elissa M. Aminoff⁵¹Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, Pennsylvania²Department of Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania³Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania⁴Machine Learning Department, Carnegie Mellon University, Pittsburgh, Pennsylvania⁵Department of Psychology, Fordham University, Bronx, New York

Correspondence

Ying Yang, Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA.

Email: ying.yang.cnb.cmu@gmail.com

Funding information

Henry L. Hillman Foundation, Grant/Award Number: Presidential Fellowship at Carnegie Mellon Univ; National Institute of Mental Health, Grant/Award Number: RO1 MH64537; National Science Foundation, Grant/Award Number: 1439237

Abstract

The human visual cortex is organized in a hierarchical manner. Although previous evidence supporting this hypothesis has been accumulated, specific details regarding the spatiotemporal information flow remain open. Here we present detailed spatiotemporal correlation profiles of neural activity with low-level and high-level features derived from an eight-layer neural network pretrained for object recognition. These correlation profiles indicate an early-to-late shift from low-level features to high-level features and from low-level regions to higher-level regions along the visual hierarchy, consistent with feedforward information flow. Additionally, we computed three sets of features from the low- and high-level features provided by the neural network: object-category-relevant low-level features (the common components between low-level and high-level features), low-level features roughly orthogonal to high-level features (the residual Layer 1 features), and unique high-level features that were roughly orthogonal to low-level features (the residual Layer 7 features). Contrasting the correlation effects of the common components and the residual Layer 1 features, we observed that the early visual cortex (EVC) exhibited a similar amount of correlation with the two feature sets early in time, but in a later time window, the EVC exhibited a higher and longer correlation effect with the common components (i.e., the low-level object-category-relevant features) than with the low-level residual features—an effect unlikely to arise from purely feedforward information flow. Overall, our results indicate that non-feedforward processes, for example, top-down influences from mental representations of categories, may facilitate differentiation between these two types of low-level features within the EVC.

KEYWORDS

neural network, nonfeedforward, spatiotemporal neural activity, visual cortex

1 | INTRODUCTION

Humans can effortlessly recognize objects and understand scenes. What computational processes in the visual cortex result in such proficiency? Decades of research has mapped the visual cortex into multiple functional brain areas, which span those defined through a topological mapping (i.e., “retinotopy”) of the visual field (Engel, Glover, & Wandell, 1997) and those defined through preferred

stimulus categories (e.g., the scene-selective “parahippocampal place area,” Epstein, Harris, Stanley, and Kanwisher (1999) and the object-selective “lateral occipital complex,” Grill-Spector et al. (1999)). More generally, the visual cortex is hypothesized to process visual inputs in a hierarchical manner (DiCarlo & Cox, 2007). At a low level, neurons in the primary visual cortex (V1) have small receptive fields and extract low-level information such as local edge orientations (Hubel & Wiesel, 1968). Next in the hierarchy, neurons in V2 are sensitive to

more complex features of the visual inputs, such as angles and junctions of edges (Ito & Komatsu, 2004), as well as textures (Freeman, Ziemba, Heeger, Simoncelli, & Movshon, 2013). This pattern continues, with regions of the visual cortex progressively processing more and more complex visual representations—ultimately resulting in high-level vision. For example, neurons in the inferior temporal cortex show selectivity to complex shapes, invariance to scales and lighting, and encode semantic information associated with the visual inputs (Tanaka, 1996). In previous literature, there is an extensive body of experimental evidence supporting this hierarchical organization, ranging from electrophysiology studies in primates to neuroimaging studies in humans (e.g., Aminoff et al. (2015); Clarke, Devereux, Randall, and Tyler (2014); Cichy, Khosla, Pantazis, Torralba, and Oliva (2016b); Yamins et al. (2014), etc).

Within the framework of a progressive visual hierarchy, as one moves from the posterior regions (e.g., V1) to the more anterior regions (e.g., category-selective regions) along the visual pathways, neurons at each level of the hierarchy receive inputs from previous levels and extract increasingly higher-level information. At the same time, we should ask, “Does information necessarily flow only in a bottom-up, feedforward direction?” Apparently not—there is also strong evidence for *feedback* anatomical connections between visual areas, that is, connections that pass information in an inverse direction with respect to the hierarchy. Similarly, there is evidence for anatomical connections skipping across levels (e.g., V4 to V1), as well as from the frontal and parietal lobes to visual areas (Felleman & Van Essen, 1991). These various connections indicate that non-feedforward dynamics are likely to exist and, moreover, that this sort of information flow may function as top-down feedback, thereby facilitating rapid visual processing and recognition (e.g., as suggested by Bar et al. (2006)).

Although there is some functional evidence for top-down information flow in the visual cortex, much of the support for feedback is neuroanatomical. As such, further studies of the spatiotemporal dynamics of visual processing are critical. For example, to address the directionality of information flow when visual inputs are processed, it is important to know not only where information is processed, but also when it is processed by one brain region relative to other brain regions. That is, one should be able to explicate the level of information extracted at a given time point from a specific brain region and compare this to information extracted from other regions at other time points. Such data would inform us regarding whether information also flows in the direction of the apparent feedback when visual inputs are processed. Moreover, if evidence for feedback is found, when and in what brain areas does such feedback occur? Details of information flow dynamics are crucial for building mathematical models of the brain and understanding how visual recognition is achieved.

To address such questions we need to better understand the neural dynamics during the perception of a rich set of visual stimuli (e.g., naturalistic, everyday visual scenes). Moreover, the ideal measurement tool should provide both good temporal and spatial resolution, as well as spatial coverage of the visual cortex. In particular, temporal resolution should be smaller than 100 ms, considering the

short latency measured in object and scene recognition (Thorpe, Fize, & Marlot, 1996). In this light, imaging methods that depend on blood oxygen level—most prominently, functional magnetic resonance imaging (fMRI)—are not well suited for our purposes. In contrast, magnetoencephalography (MEG) offers millisecond-level temporal resolution, enabling the measurement of rapid neural processing in object, and scene recognition. Importantly, MEG also provides spatial coverage of the visual system; although MEG can not provide millimeter-level spatial resolution as fMRI (Sejnowski, Churchland, & Movshon, 2014), using source localization techniques (e.g., the techniques in Dale et al. (2000); Hamalainen, Hari, Ilmoniemi, Knuutila, and Lounasmaa (1993); Yang, Tarr, and Kass (2014)), MEG is likely to differentiate neural activity from low-level and high-level portions of the visual cortex. For example, Hedrich, Pellegrino, Kobayashi, Lina, and Grova (2017) recently demonstrated that, with high-density (more than 256) sensors, the median localization error of some common source localization methods were 20 to 40 mm. In this context, MEG is capable of supporting a relatively detailed spatiotemporal profile of neural activity. As such, we have adopted MEG as a promising neuroimaging method for noninvasively studying neural responses to naturalistic visual inputs and addressing the question—*what is the form of information extracted at different temporal stages and different spatial locations during scene and object perception?*

To explore the level of information representation (e.g., edges vs. full shapes vs. meaningful object categories) in the spatiotemporal dynamics of neural processing, one can regress measured spatiotemporal neural activity during a visual task against different candidate feature sets that characterize different levels of information (Nestor, Vettel, & Tarr, 2008). However, defining the candidate features is nontrivial. Building on a tradition of applying computer vision models to the understanding of biological vision (Leeds, Seibert, Pyles, & Tarr, 2013), many different groups have recently used a type of artificial-neural-network model, known as convolutional neural networks (“CNNs”), to study human visual processing (Yamins & DiCarlo, 2016). What makes CNNs particularly attractive as “proxy models” for studying biological vision is their extremely high performance in categorizing both naturalistic objects and scenes, as well as their demonstrated ability to learn task-relevant, diagnostic features across very large image sets (Krizhevsky, Sutskever, & Hinton, 2012). Interestingly, CNNs (and related “deep” networks) were inspired by feedforward processing as embodied in the hypothesized hierarchical structure of the primate visual cortex (LeCun, Bengio, & Hinton, 2015). In particular, CNNs typically incorporate multiple layers with units of increasing receptive field sizes as one moves up along the hierarchy; the first layer takes raw image inputs and the last layer outputs object or scene category labels. The connections between layers are optimized by minimizing labeling error on a large amount of image data. After such optimization, the layered structure of the network inherently provides an operational definition of low-level to high-level representation of task-relevant visual information (Zeiler & Fergus, 2014). Moreover, as alluded to earlier, features extracted from CNNs have been found to share significant representational similarity with the neural representations of objects and scenes at both low levels, as exemplified by

early visual regions, and high levels, as exemplified by category-selective brain regions (Cichy, Khosla, Pantazis, & Oliva, 2017; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016c; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014).

In this context, we used CNNs to explicate the level of visual representation within a spatiotemporal context. To the degree that we can use this approach to associate low- and high-level visual features with specific processing time points, we can develop a more detailed account of feedforward versus feedback communication within the visual cortex. More specifically, we regressed the neural responses—as measured by MEG—at different cortical locations and time points against low- and high-level features from a high-performing CNN. Critically, this regression analysis explores how well neural responses are explained by particular feature sets, yielding a spatial-temporal profile of statistical dependence between neural activity and different levels of CNN features. Here, we used linear models for the regression. Indeed, the spatiotemporal profile of linear dependence, which we refer to as the “correlation profile,” is the key result we present in this article.

Our choice of linear models is motivated by two different perspectives. First, from a scientific perspective, as proposed by DiCarlo and Cox (2007), along the hierarchical organization of the visual cortex, neural substrates appear to perform a series of transformations, such that the representation of different categories of objects or scenes become progressively more separable (e.g., by linear classifiers). In this sense, in relating the neural activity of a given brain region to a given feature set, it is reasonable to use low-capacity linear models, essentially only using a “minimum” transformation; high-capacity nonlinear models may add additional transformations to the given features, which may correspond to the computation by other parts in the brain at a higher level of processing. Second, from a statistical perspective, although we could adopt a high-capacity regression model to capture nonlinear statistical dependence, assuming unlimited number of observations of neural responses to naturalistic stimuli, in practice, the number of observations we can actually acquire is limited by our experimental methods. Using linear regression can restrict model complexity and avoid over fitting; thus they are better suited for this study.

Several recent MEG studies have likewise related neural activity to CNN or other computer vision derived features. However, in contrast to our present study, these studies focused on recognizing isolated objects on blank backgrounds (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016c; Clarke et al., 2014), or on specific properties of scenes, restricting the input to a small set of scene stimuli (Cichy, Khosla, Pantazis, & Oliva, 2017). For the most part, these studies also focused only on temporal patterns in MEG, but not on spatial patterns (although see Clarke et al. (2014)). To address what we view as limitations, we recorded neural responses to a relatively large number of *everyday, naturalistic scenes* and applied *source-space regression analysis*, aiming to obtain detailed correlation profiles between spatiotemporal neural activity and low- and high-level CNN features.

Beyond these extensions of previous work, our present study is distinguished by our implementation of a unique decomposition of low-level and high-level CNN features. The high-level features are

nonlinear transforms of the low-level features; features at different levels may share common linear components. The existence of these common components renders it difficult to compare correlations between neural data and the features at different levels (as has been done previously); that is, correlations with the low- and high-level features may be driven by common components. In response to this challenge, we present a new approach, which extracts the common components in the low- and high-level CNN features and decomposes them into three groups:

1. *Common components* between the low-level and high-level features, or the subspace of low-level features that are relevant to high-level features.
2. *Residuals of low-level features*. These are the low-level features that are roughly orthogonal to the high-level features, or the remaining features after regressing out the common components.
3. *Residuals of high-level features*. These are the high-level features that are roughly orthogonal to the low-level features, or the remaining features after regressing out the common components.

By comparing the spatiotemporal correlation profiles of these three groups, we are able to observe dynamics for feedforward processing in the visual cortex. Additionally, we see evidence for non-feedforward processing, which we hold may reflect top-down feedback that is deployed as part of naturalistic scene processing (Bar et al., 2006).

2 | MATERIALS AND METHODS

2.1 | Participants

Eighteen healthy adults (8 females/10 males, 1 left-handed/17 right-handed, age range 18–30) participated in both the MEG and the fMRI sessions. All participants provided written informed consent and were financially compensated for their participation. All procedures followed the principles in the Declaration of Helsinki and were approved by the Institutional Review Boards of Carnegie Mellon University and the University of Pittsburgh.

2.2 | Stimuli

The stimulus set used in the MEG sessions included color images (photographs) of 181 scene categories (e.g., “airport,” “beach,” “coffee shop,” etc.). There were two exemplar images for each category, resulting in 362 images in total. These images were obtained from the data set used by the “Never Ending Imaging Learner” (NEIL, Chen, Shrivastava, and Gupta (2013), www.neil-kb.com), which were originally obtained from the Internet. The scene images varied in aspect ratio with the longest dimension (either the width or the height) set to 500 pixels. The images were placed in the center of a 600 × 600 bounding box filled with a gray value = 135 (of 255). Figure 1 shows two example stimuli. The images (with the gray bounding boxes) were presented at a visual angle of about 10 by 10°.

FIGURE 1 Illustration of the trial structure in the MEG session [Color figure can be viewed at wileyonlinelibrary.com]

In the fMRI session (see below), functional localizer scans in which a separate set of images were presented, were also included, in order to independently define the object- and scene-selective regions in the visual cortex. The localizer images included 60 color images in each of the three conditions: scenes, objects, and phase-scrambled pictures of the scenes. The scene images included outdoor and indoor scenes, which did not overlap with the 362 scene images discussed above. The particular objects used were weak contextual objects as described in Bar and Aminoff (2003). The phase-scrambled pictures served as control stimuli for scenes; they were generated by running a Fourier transform of each scene image, scrambling the phases, and then performing an inverse Fourier transform back into the pixel space. The localizer images were presented at a visual angle of 5.5 by 5.5° .

2.3 | Experimental procedure

The experimental procedures in the MEG and MRI sessions were implemented using **Psychtoolbox 3** (<http://psychtoolbox.org/>) in **MATLAB**. The trial structure in the MEG sessions is illustrated in Figure 1. Before the stimulus presentation, the participants were asked to fixate their eyes on a white “+” symbol (spanning 80 pixels or 1.3° at the center of the screen). The gray value of the screen was 180 (of 255). We term this screen “the fixation screen” hereafter. Following the fixation screen, one stimulus image was presented at the center of the screen for 200 ms, which was short enough to reduce artifacts due to saccades during stimulus presentation. The stimulus image was then followed by the fixation screen lasting for a random duration until the onset of the next stimulus. This duration was uniformly sampled from a (1,600, 2,100) ms interval independently in each trial. Participants performed a “one-back” task — that is, they responded, by pressing a button on a glove pad using the right index finger, when the current stimulus was a repetition of the stimulus from the immediately preceding trial. Participants were given 1,500 ms to respond after the stimulus onset.

Each MEG session included 6 to 12 runs; each run included 181 images, and every two runs covered all 362 images. In each run, 10% of the images were immediately repeated once, and only the data corresponding to the first presentation of each image were analyzed. In total, there were three to six repetitions of each image for most of

the participants in the MEG session. For one participant, due to a problem with acquisition hardware, only two repetitions were presented for some images, and four repetitions were presented for the remaining images.

Each fMRI session included functional localizer scans to independently define scene- and object-selective regions in the visual cortex; during the scans, participants viewed images of scenes, weak-contextual objects and phase-scrambled scenes while doing a one-back task—pressing a button on a glove pad with the right index finger when there was an immediate image repetition. Most participants went through two runs of the localizer scans; however, two participants went through one run due to time limitations. Each run started and ended with a 12-s time window, during which a black fixation cross (“+”) was presented on a gray background. Between the starting and ending fixation windows, there were twelve 14 to 16-s blocks, each of which presented stimuli in one condition (scenes, weak-contextual objects, or phase-scrambled scenes). There were four blocks per condition and three conditions in total; there was also a 10-s fixation time window between two consecutive blocks. In each block, 14 stimuli were presented in a row, with an 800 ms presentation duration and a 200 ms inter-stimulus interval, with the exception that the first stimulus in each of the block other than the first block was presented for 2,800 ms, yielding 16-s blocks. Note that the longer presentation was due to a timing issue in the customized image presentation program. However, we are doubtful that this variation changed our results in defining the object/scene-selective regions. In particular, we relied on a block design in which the main effect is the difference of neural responses for different blocked conditions. As such, these results are expected to be robust to small variations in the presentation duration of individual stimuli. Among the 14 stimuli, 12 were unique images, and 2 were immediate repetitions of the previous image in order to allow for positive responses in the one-back task.

2.4 | Data acquisition

2.4.1 | Magnetoencephalography

MEG data were collected using a 306-channel whole-head MEG system (Elekta Neuromag, Helsinki, Finland) at the Brain Mapping Center

at the University of Pittsburgh. The MEG system had 102 triplets, each consisting of a magnetometer and two perpendicular gradiometers. The recordings were acquired at 1 kHz, high-pass filtered at 0.1 Hz and low-pass filtered at 330 Hz. Four head position indicator (HPI) coils were placed on the scalp of each participant to record the position of the head in relation to the MEG helmet. Empty room MEG recordings were collected in the same session, and used to estimate the covariance matrix of the sensor noise. Approximately 100 points describing the shape of the head and the coordinates of the HPI coils on the scalp were collected using a digitization device; these coordinates were later used in aligning the head position in the MEG session with the structural MRI scan.

Electrooculogram (EOG) was monitored by recording electric potentials above and below one eye and lateral to both eyes; electrocardiography (ECG) was recorded by placing additional electrodes under the clavicles. The EOG and ECG recordings captured eye blinks and heartbeats, allowing these artifacts to be removed from the MEG recordings during data preprocessing.

2.4.2 | Magnetic resonance imaging

MRI data were collected on a 3 T Siemens Verio MR scanner at the Scientific Imaging and Brain Research Center at Carnegie Mellon University using a 32-channel head coil. First, for each participant, a high-resolution structural MRI scan was acquired (T1-weighted MPRAGE sequence, $1 \times 1 \times 1 \text{ mm}^3$, 176 sagittal slices, TR = 2,300 ms, TE = 1970 ms, flip angle = 9° , GRAPPA = 2, field of view = 256). Second, functional MRI (fMRI) data were collected for the functional localizer (T2*-weighted echo-planar imaging multiband pulse sequence, 69 slices aligned to the AC/PC, in-plane resolution $2 \times 2 \text{ mm}^2$, 2 mm slice thickness, no gap, TR = 2000 ms, TE = 30 ms, flip angle = 79° , multiband acceleration factor = 3, field of view 192 mm, phase encoding direction A $\gg$ P, ascending acquisition). Third, a fieldmap scan was acquired to correct for distortion effects using a similar slice prescription as the echo-planar imaging scans (69 slices aligned to the AC/PC, in-plane resolution $2 \times 2 \text{ mm}^2$, 2 mm slice thickness, no gap, TR = 724 ms, TE1 = 5 ms; TE2 = 7.46 ms, flip angle = 70° , field of view 192 mm, phase encoding direction A $\gg$ P, interleaved acquisition).

2.5 | Preprocessing of MEG data

Preprocessing of the raw MEG data were accomplished via the following pipeline. All steps in this pipeline were implemented using the **MNE-python** package (Gramfort et al., 2014) in **Python**.

1. *Filtering.* The raw recordings (including the MEG empty room recordings) were filtered with a 1–110 Hz bandpass filter, which removed low-frequency drifts and high frequency noise (such as the oscillations generated by the head position indicator coils during head tracking). Recordings were further filtered by a notch filter centered at 60 Hz intended to remove the power line interference.
2. *Removing artifacts due to eye blinks and heartbeats.* Independent component analysis (ICA) was used to decompose the

306-dimensional recordings into more than 100 independent components. For the majority of the participants, two components that were highly correlated with EOG and one to two components that were highly correlated with ECG were manually identified by the authors; these components, which were likely due to blink and heartbeats, were excluded in reconstructing the data using the independent components.

3. *Obtaining trial-by-trial data.* Trial-by-trial recordings (also referred as “epochs”) were obtained by segmenting the data from –100 to 900 ms with respect to the “trigger” onset (the “stimulus onset” as recorded in the acquisition system, defined as 0 ms). For each trial and each channel, the mean across time points in the baseline window (–100 to 0 ms) was subtracted from the recording at each time point. Note that the timing here was recorded by the data acquisition device; the image presentation device had a measured delay of 40 ms according to a photosensor placed on the screen. As such, to correctly align the data, we shifted all time points backward by 40 ms; in this sense, the baseline time window became –140 to –40 ms.

A signal space projection (SSP) was applied to all epochs. The SSP constructed a low-dimensional linear subspace characterizing the empty room noise (via principal component analysis [PCA]), and removed the projection onto this subspace from the experimental MEG recordings. As such, only those neural signals orthogonal to the principal components of empty room noise remained. Two principal components (PCs) for 204 gradiometers and three principal components for 102 magnetometers were used to create a span of the noise space from empty room recordings. Across participants, the mean proportion of variance in the experimental data explained by these PCs were 2.5%, with a standard deviation (SD) of 1.3%.

4. *Obtaining averaged neural responses to each image for each participant.* To reduce the computational cost for analyses discussed below, we down-sampled the trial-by-trial data to 100 Hz (i.e., the sampling rate = 100 Hz). In addition, those trials corresponding to the second presentation in immediate image repetitions may have had lower signal strength due to adaptation; therefore, they were removed from further analyses. To remove outlier trials that had extreme large variations in each session for each participant, we computed the difference between the maximum and the minimum of the recordings for each channel in each trial, and discarded those trials where the difference was larger than 15 standard deviations plus the mean across all trials for at least one channel. On average across participants, 6.7 trials of 1,086 ~ 2,172 trials ($362 \times [3 \text{ to } 6]$) were removed in this outlier detection procedure. The removed outliers were $0.3 \pm 0.2\%$ of the total number of trials (mean $\pm$ SD across participants). For each image and each participant, there was at least one trial left after removing outliers. Finally, the data in the remaining trials that corresponded to the same image were averaged within each session for each participant.
5. *Regressing out neural data explained by nuisance covariates.* Although our stimulus images were all displayed in the same 600×600 pixel boxes, the widths and heights of the images themselves varied. Such nuisance covariates are irrelevant to the image contents, but could explain a significant amount of variance

in the MEG recordings. To remove such spurious effects, we regressed the MEG data against four covariates—image width, image height, area (width $\times$ height) and aspect ratio (width/height)—respectively at each time point for each sensor and each participant. An all-one column was added to the regressors to remove the mean response across all images as well. The residuals were then retained as new sensor data to be analyzed.

2.6 | Forward modeling

For each participant, based on the T1-weighted structural MRI scan, the outer skin surface of the scalp and the inner and outer surfaces of the skull were computed using the watershed algorithm (Ségonne et al., 2004) implemented in the **FreeSurfer** software (Fischl et al., 2002) and the **MNE-C** software (Gramfort et al., 2014 https://martinos.org/mne/dev/install_mne_c.html). Additionally, the cortical surfaces that segmented the gray matter and the white matter were also computed using **FreeSurfer**. The source space was defined as about 8,000 distributed “dipoles” (or source points) that pointed perpendicularly to the cortical surface of both hemispheres. The average spacing between source points was 4.9 mm, yielding 24 mm² of cortical surface area per source point. Source points that were within 2.5 mm of the inner skull surface were excluded.

The digitized points that described the shape of the head in the MEG session were used for co-registration with the structural MRI. In the co-registration, we solved for a rigid-body transformation that minimized the sum of squared distances from the digitized points to the scalp surface, using an interface implemented in **MNE-C** (Gramfort et al., 2014). Note that the optimization problem was not necessarily convex, therefore no global minimum could be guaranteed. Yet, by manually adjusting initial values, the solution for each participant appeared to be good in visual inspection.

For each participant, the forward matrix for each run in the MEG session was computed using the boundary element model implemented in **MNE-C**, after transforming the MEG sensor locations into the structural MRI space, based on the alignment in the co-registration step above. The forward matrices across all runs were averaged for each participant.

2.7 | Source localization and source-space regression analysis

The dynamic statistical parametric mapping (dSPM) (Dale et al., 2000) source localization method implemented in **MNE-python** was used to obtain unit-less (i.e., standardized) source current dipole estimates. This method estimates a linear projection that projects the MEG sensor data into the source space and then normalizes the projections with estimated standard deviations. Since there were more source points on the cortical surface than the number of sensors, a penalization parameter specifying a zero-mean independently distributed Gaussian prior with a constant variance of the source estimates was used. The penalization parameter was set to 1.0; it was selected based on suggested setting in **MNE-python**, which is empirical in nature. The **MNE-python** software

suggested using signal-to-noise ratio (SNR) to select the penalty, setting the penalty = $1/\text{SNR}^2$. We assumed SNR = 1.0 in the analysis. We also ran the analysis with other penalty values and found the results were qualitatively the same when the penalization parameter was set to 0.1 (SNR $\approx$ 3.16) and 10 (SNR $\approx$ 0.316). The noise covariance matrix was estimated from sensor recordings within the baseline time windows (−140 to −90 ms) for each participant.

To characterize how much the spatiotemporal neural activity was correlated with a given set of CNN features, we regressed the neural responses in the source space against the CNN-derived features of each image. Following obtaining the dSPM solutions for each image, an ordinary least square regression was run for each participant, for each time point and each source point. The coefficient of determination (or R-squared), indicating the proportion of variance explained by the regressors (i.e., CNN features) was used as the summarizing statistic to characterize the correlation between neural responses and a given set of CNN features.

2.8 | Preprocessing of fMRI data and definition of regions of interest

The fMRI localizer data were preprocessed in **SPM12** (<http://www.fil.ion.ucl.ac.uk/spm/software/spm12/>). The preprocessing included an unwarp transform to correct for geometric distortions using the fieldmap scan, a frame-by-frame transform to correct for head motion combined with a transform to align with the structural MRI, and finally spatial smoothing with an isotropic Gaussian kernel (where the full width at half maximum was 4 mm). The data in all localizer runs for each participant were concatenated, high pass filtered with a cut-off frequency at 0.0078125 Hz (a 128-second period), and then analyzed using a “general linear model” in a block design. In this model, the time series at each voxel was linearly regressed against a design matrix, which included, in separate columns, the predefined canonical hemodynamic response function convolved with the boxcar-like indicators of blocks for each stimulus condition (scenes, weak contextual objects, and phase scrambled scenes). The design matrix also included extra columns corresponding to nuisance covariates (e.g., the time series of parameters in motion correction). An autoregressive model of order 1 (AR[1]) was used to account for the temporal correlations in the residuals.

Scene/object selective regions were defined using the **MarsBaR** toolbox (<http://marsbar.sourceforge.net/index.html>). For any voxel, let β_{scene} denote the regression coefficient for the scene condition, β_{object} for the weak-contextual object condition, and finally β_{scramble} for the phase-scrambled scene condition. The t -statistics of the difference ($\beta_{\text{scene}} - 1/2(\beta_{\text{object}} + \beta_{\text{scramble}})$) was computed as the estimated difference divided by the estimated SD of the difference. Then the voxels where the t -statistics was above a threshold were selected as scene-selective voxels. A customized threshold was set for each participant, such that clusters of contiguous voxels above the threshold were identified within or in the proximity of the parahippocampal gyrus, the retrosplenial cortex, and the transverse occipital sulcus (TOS) in each hemisphere. These clusters were labeled as the scene-selective regions of interest (ROIs), which were the parahippocampal

place area (PPA), the retrosplenial complex (RSC) and the occipital place area, also labeled as the TOS area. For the majority of the participants, the threshold was equal to the value where the family-wise error rate was controlled at 0.05. For some individuals, if the threshold was too stringent, we relaxed the threshold to a smaller value. Similarly, object-selective clusters in the lateral occipital cortex and the fusiform area (i.e., voxels in the lateral occipital complex or LOC) were also identified in both hemispheres for each participant respectively, where the difference of interest was $\beta_{\text{object}} - \beta_{\text{scramble}}$.

The **SUMA** software (<https://afni.nimh.nih.gov/Suma>, Cox, 1996) was used to project these clusters of voxels to sets of vertices on the cortical surfaces, so that they could be labeled in the source space in MEG. The mapped sets of vertices were then manually examined, and the ones that had fuzzy boundaries or that were anatomically off were corrected. Each remaining contiguous set was defined as one region of interest (ROI). The ROIs that did not cover at least 10 source points (in the MEG source space) were dilated until they covered 10 source points. Finally, the vertices in the LOC that were also in one of the scene-selective ROIs (PPA, RSC, or TOS) were removed from the LOC. Additionally, some ROIs were defined based on the parcellation of the structural MRI by **FreeSurfer**. These ROIs included the left and right pericalcarine areas that covered the early visual cortex (EVC)—the pericalcarine areas included mostly V1 but might also include some part of V2. Finally, each pair of the corresponding bilateral regions were merged into one ROI, resulting in the bilateral PPA, RSC, TOS, LOC, and EVC. See Figure S1 in Appendix for the locations of the ROIs for each participant.

2.9 | Extracting features of the images

We used a convolutional neural network called **Alexnet** (Krizhevsky et al., 2012) implemented in the **Caffe** software (Jia et al., 2014) to extract features. This CNN was trained to classify images into 1,000 object categories with 1.2×10^6 training samples and 5×10^4 validation samples in the **ImageNet** database (Deng et al., 2009; Russakovsky et al., 2015). The first five layers had convolutional units. Each unit in these layers applied a dot product of a “kernel” weight matrix with the inputs within a receptive field—for example, in Layer 1, each unit’s receptive field was 11×11 pixels of the RGB channels of the raw image. Within each of the convolutional layers, there were a number of sub-layers (or channels); for example, Layer 1 had 96 sub-layers. Units in each sub-layer shared the same “kernel” weight matrix across all spatial locations; therefore, in such a convolutional architecture, the number of parameters was much smaller than that of a fully connected neural network with the same number of units, yielding a more tractable model to train. After the convolution operation (dot product), each unit then applied a rectified linear function $f(x) = \max(0, x)$ on the dot product to generate the output. In Layers 1, 2, and 5, an additional normalization step was applied, where in each location, the output of the unit in each sub-layer was normalized by a function of the sum of squares of the responses in its neighbor sub-layers, including itself; a max-pooling operation was also added after the normalization in these layers (for details, see Krizhevsky

et al., 2012). Layers 6 and 7 were fully connected layers, each consisting of 4,096 units, and Layer 8 was the final output layer with 1,000 units, corresponding to the 1,000 object categories. We downloaded a pretrained version of **Alexnet** from the **Caffe** “model zoo,” and re-sized our 600×600 images (including the gray box) to the input size required by **Alexnet**. For each image, we collected responses for all the units in each layer and concatenated them into a vector. For Layers 1, 2, and 5, we used the responses before normalization and max-pooling.

Note that the eight-layer structure of **Alexnet** is feedforward; it naturally provides a progressive shift from low-level to high-level features. Here we chose Layer 1 and Layer 7 as representatives of low-level and high-level (object-category-related) features, respectively. Consistent with this approach, Layer 1 had the smallest receptive field sizes and the convolutional “kernel” weight matrices were similar to 2-D Gaussian functions and Gabor filters (see the visualization in Krizhevsky et al., 2012). In contrast, Layer 7 was the last fully connected hidden layer before the output layer and was expected to represent task-relevant semantic information about the images. We did not include Layer 8 features. This decision was based on the observation that the correlation effect between the neural activity and Layer 8 features was much weaker than with other layers (see Figure 3 below). Layer 8 was the output layer, and the output 1,000 category labels were defined according to WordNet (Deng et al., 2009; Fellbaum, 1998), instead of a data-driven manner using brain activity. Therefore, these categories may not necessarily be the best ones to represent the organization of objects in the visual cortex.

As mentioned earlier, nuisance covariates due to the various widths and heights of the stimulus images might also have been correlated with the extracted CNN features. Therefore, we regressed out the width, height, area (width $\times$ height) and aspect ratio (width/height) from the features extracted from each unit across the union of the stimulus image set and additional image set (see below). An all-one column was included in the regression, which removed the mean across all images.

In addition to the **AlexNet** features, we also extracted a set of simpler low-level features—the **local contrast** features—in the following way. For each image re-sized to the input size of **Alexnet**, the patch within each 11×11 receptive field for Layer 1 was converted to gray values (by averaging the values of the RGB channels), and the contrast in the patch was defined as $(x_{\text{max}} - x_{\text{min}}) / (x_{\text{max}} + x_{\text{min}})$, where x_{min} and x_{max} were the minimum and maximum of the gray values in the patch. The contrast values were concatenated across all receptive fields, yielding a $55 \times 55 = 3,025$ -dimensional vector for each image. Linear projections onto the nuisance covariates related to the widths and heights of the images were removed.

In a preliminary analysis (not included here but see Table 5.1 in Chapter 5 of Yang, 2017), we observed that the common linear components between low-level and high-level layers in **AlexNet** are correlated with **local contrast**, suggesting that there is some potential for local contrast alone to naturally elicit significant neural responses in the visual cortex. To account for this possible confounding effect due to **local contrast**, we regressed out the first 160 principal components of the **local contrast** features

FIGURE 2 The canonical correlation analysis (CCA) procedure to identify the common and residual features of Layer 1 and Layer 7. Note that 1,086 extra images of the same scene categories were introduced. At the top of the figure, the Layer 1 (blue) and Layer 7 features (orange) were obtained for the extra images and the stimulus images, and their dimensions were reduced. Afterwards, the lower-left light gray panel was run first, for selecting the value of p_c and learning the CCA projections W_1 and W_2 . At last, the darker gray panel on the lower-right was run, where the learned W_1 and W_2 were applied to the stimulus images, the common features were obtained from the canonical correlation components and the residual layers were obtained after regressing out the common features [Color figure can be viewed at wileyonlinelibrary.com]

(explaining 90% of the variance in the local contrast features) from the intact features in all layers in **AlexNet**, and took the residuals as new features of interest. We term these new features **local-contrast-reduced features** hereafter.

2.10 | Identifying the common and residual feature sets

Within **AlexNet**, as the input data progresses from Layer 1 to Layer 7, the units in each layer nonlinearly transform the raw pixel inputs to informative task-related features (where the task **AlexNet** was trained on was object categorization). In this context and as discussed earlier, Layer 1 features of **AlexNet** were identified as low-level features, while Layer 7 features were identified as high-level features. However, this does not mean that the features in Layer 1 and Layer 7 are completely orthogonal to each other. More specifically, there is likely to be some linear dependence between the low-level Layer 1 feature set and the higher-level Layer 7 feature set. As such, correlating neural data separately with both Layer 1 and Layer 7 does not allow us to infer the degree to which these different correlations are due to the common linear components between the two layers. To address this ambiguity, we used canonical correlation analysis (CCA) to extract the *common linear components* between Layer 1 and Layer 7. In particular, for two multivariate variables, CCA learns a linear projection for each variable, such that the correlation between the projected results from the two variables is maximized. Similar to PCA, we can learn more than one pair of linear projections, and the projections for each original variable are orthogonal to each other.

Figure 2 illustrates the mathematical procedure of the CCA. Because the CNN features were high dimensional, we included a set of

additional images (“the extra images”) to boost our ability of analyzing the features. These extra images were from the same 181 scene categories in the same image data set, including six exemplars per category, different from the stimulus images presented in the MEG experiment. These extra images ($6 \times 181 = 1,086$ images in total) had similar sizes (longest side = 500 pixels) as the stimulus images, and were also centered in the same 600×600 gray boxes. The CNN features of these images were obtained in the same way as the stimulus images. The local contrast features for each additional image were obtained as well; across both the stimulus images and the additional images, a PCA was applied on these local contrast features, and 160 of these components, which explained 90% of the variance in the local contrast features, were regressed out from the features in each layer across all the images, resulting in the **local-contrast-reduced features**, which were then used for the analysis detailed below.

Let matrices X_1 ($q \times p_1$) and X_2 ($q \times p_2$) denote two sets of features of the same q images, extracted from p_1 units in Layer 1 and p_2 units in Layer 7 of **Alexnet**. We assumed that the rows in X_1 and X_2 had zero means, which was empirically satisfied by subtracting the sample mean across q images. Because there were more units in the layers of **AlexNet** than the number of images ($q < p_1, p_2$), we first used PCA to reduce the dimensions of both feature sets to $p = 362 < p_1, p_2$. With the assumption that both X_1 and X_2 had zero mean, the PCA was implemented using singular value decomposition,

$$X_1 \approx U_1 D_1 V_1^T, \quad X_2 \approx U_2 D_2 V_2^T$$

$q \times p_1 \quad q \times p \quad p \times p \quad p \times p_1 \quad q \times p_2 \quad q \times p \quad p \times p \quad p \times p_2$

where D_1 and D_2 were diagonal matrices, and U_1 , U_2 , V_1 and V_2 had orthonormal columns. The notation V^T denotes the transpose of matrix V .

FIGURE 3 Proportion of variance explained by the first 10 principle components of the **AlexNet** features in each layer, averaged across sensors. Each color corresponds to one layer, and the confidence intervals were shown in transparent bands. On the left, the features were intact before removing the local-contrast components; on the right, the features were the *local-contrast-reduced* features. The lines below the plots indicate in which time windows the proportion of variance explained was significantly higher than that of the baseline window before the stimulus onset. (a) Intact features (b) local-contrast-reduced features [Color figure can be viewed at wileyonlinelibrary.com]

We used the projections of $\mathbf{X}_1$ and $\mathbf{X}_2$ onto the p orthogonal dimensions $\mathbf{X}_1 = \mathbf{U}_1 \mathbf{D}_1$ and $\mathbf{X}_2 = \mathbf{U}_2 \mathbf{D}_2$ in the CCA, where linear weights for the projections $\mathbf{W}_1$ and $\mathbf{W}_2$ (both $p \times p$) were obtained as follows.

$$\begin{aligned} & \arg \max_{\mathbf{W}_1[:,i], \mathbf{W}_2[:,i]} \text{cov}(\mathbf{X}_1^\dagger[:,i], \mathbf{X}_2^\dagger[:,i]) \text{ where } \mathbf{X}_1^\dagger = \tilde{\mathbf{X}}_1 \mathbf{W}_1, \mathbf{X}_2^\dagger = \tilde{\mathbf{X}}_2 \mathbf{W}_2 \\ & \text{subject to } \text{cov}(\mathbf{X}_1^\dagger[:,i], \mathbf{X}_1^\dagger[:,j]) = 0 \text{ if } i \neq j \\ & \text{cov}(\mathbf{X}_2^\dagger[:,i], \mathbf{X}_2^\dagger[:,j]) = 0 \text{ if } i \neq j \\ & \text{var}(\mathbf{X}_1^\dagger[:,i]) = \text{var}(\mathbf{X}_2^\dagger[:,i]) = 1 \\ & i, j = 1, \dots, p \end{aligned}$$

The i th columns of $\mathbf{W}_1$ and $\mathbf{W}_2$, ($\mathbf{W}_1[:,i]$ and $\mathbf{W}_2[:,i]$) were the weights that linearly combined columns in $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2$, such that the combinations $\mathbf{X}_1^\dagger[:,i]$ and $\mathbf{X}_2^\dagger[:,i]$ had the highest correlation. Different columns in $\mathbf{W}_1$ and $\mathbf{W}_2$ projected $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2$ onto orthogonal components. This optimization problem was solved using the `canoncorr` function in **MATLAB**.

The solution yielded p orthonormal components in $\mathbf{X}_1^\dagger$ and $\mathbf{X}_2^\dagger$, where the correlations between the corresponding components in the two feature sets decreased from the first to the p th components. To determine how many components to include in further analyses, we used cross validation error in predicting one feature set from the other as a measurement of goodness. Assume we used the top p_c components in the prediction, for each $i = 1, \dots, p_c$, we learned a linear regression $\mathbf{X}_2^\dagger[:,i] \approx a_i \mathbf{X}_1^\dagger[:,i] + b_i$. Given new observations $\tilde{\mathbf{X}}_1^{\text{new}}$, we predicted $\tilde{\mathbf{X}}_2^{\text{new}}$ using the following pipeline.

$$\tilde{\mathbf{X}}_1^{\text{new}} \xrightarrow{\mathbf{X}_1^\dagger = \tilde{\mathbf{X}}_1 \mathbf{W}_1} (\mathbf{X}_1^\dagger)^{\text{new}} \xrightarrow{[:,1:p_c] \text{ linear mapping}} \text{prediction } \tilde{\mathbf{X}}_2^\dagger[:,1:p_c] \xrightarrow{\text{least square}} \text{prediction } \hat{\tilde{\mathbf{X}}}_2$$

The last arrow above involved solving the following least square problem

$$\arg \min_{\hat{\tilde{\mathbf{X}}}_2} \left\| \hat{\tilde{\mathbf{X}}}_2 \mathbf{W}_2[:,1:p_c] - \tilde{\mathbf{X}}_2^\dagger[:,1:p_c] \right\|_F^2$$

where $\|\cdot\|_F$ was the Frobenius norm. Setting this objective function's gradient to zero, we got

$$\begin{aligned} \hat{\tilde{\mathbf{X}}}_2 &= \left(\tilde{\mathbf{X}}_2^\dagger[:,1:p_c] \mathbf{W}_2[:,1:p_c]^T \right) \left(\mathbf{W}_2[:,1:p_c] (\mathbf{W}_2[:,1:p_c])^T \right)^{-1} \\ &= \tilde{\mathbf{X}}_2^\dagger[:,1:p_c] \mathbf{V}_{\mathbf{W}_2} \text{diag}(1/\mathbf{D}_{\mathbf{W}_2}) \mathbf{U}_{\mathbf{W}_2}^T \end{aligned}$$

where $\mathbf{U}_{\mathbf{W}_2}$, $\mathbf{D}_{\mathbf{W}_2}$, and $\mathbf{V}_{\mathbf{W}_2}^T$ were obtained from the singular value decomposition of $\mathbf{W}_2[:,1:p_c]$. The prediction error was quantified as the squared Frobenius norm of the difference between the prediction and the true $\tilde{\mathbf{X}}_2^{\text{new}}$, divided by the squared Frobenius norm of the true value (i.e., the error was $\left\| \tilde{\mathbf{X}}_2 - \tilde{\mathbf{X}}_2^{\text{new}} \right\|_F^2 / \left\| \tilde{\mathbf{X}}_2^{\text{new}} \right\|_F^2$). Similarly, a symmetric procedure could be applied to predict $\mathbf{X}_1$ from $\mathbf{X}_2$.

Noticing that selecting p_c using only the 362 stimulus images would be susceptible to overfitting, we used the 1,086 extra images for this purpose (see Figure 2). We obtained $\mathbf{X}_1$ and $\mathbf{X}_2$ from the union of the

extra images and the stimulus images, and then we ran a leave-one-exemplar-out-in-each-category (six-fold) cross validation only on the rows corresponding to the extra images, predicting $\mathbf{X}_1$ from $\mathbf{X}_2$ and vice versa. The cross validation error was computed for different values of $p_c = 1, 2, 3, \dots, 20$; we selected $p_c = 6$, the largest integer for which the cross-validated error was smaller than chance. With $p_c = 6$ the correlation between corresponding dimensions ranged from 0.8 to 0.5.

After selecting the best p_c , we applied $\mathbf{W}_1[:, 1:p_c]$ and $\mathbf{W}_2[:, 1:p_c]$ that were learned from the rows corresponding to the extra images ($\tilde{\mathbf{X}}_1^{\text{extra}}$ and $\tilde{\mathbf{X}}_2^{\text{extra}}$ in Figure 2) on the rows corresponding to the stimulus images ($\tilde{\mathbf{X}}_1^{\text{stim}}$ and $\tilde{\mathbf{X}}_2^{\text{stim}}$), obtaining the projections $(\tilde{\mathbf{X}}_1^{\text{stim}})^{\text{stim}}[:, 1:p_c]$ and $(\tilde{\mathbf{X}}_2^{\text{stim}})^{\text{stim}}[:, 1:p_c]$. Let $\mathbf{X}_{1,2}^{\dagger}$ ($362 \times 2p_c$) be the union of the columns in $(\tilde{\mathbf{X}}_1^{\text{stim}})^{\text{stim}}[:, 1:p_c]$ and $(\tilde{\mathbf{X}}_2^{\text{stim}})^{\text{stim}}[:, 1:p_c]$; then there were p_c correlated pairs in these columns. We reduced the dimension to p_c using PCA (singular value decomposition assuming zero mean), obtaining

$$\mathbf{X}_{1,2}^{\dagger} \approx \mathbf{U}_{1,2} \mathbf{D}_{1,2} \mathbf{V}_{1,2}^T,$$

$362 \times 2p_c \quad 362 \times p_c \quad p_c \times p_c \quad p_c \times 2p_c$

where we call $\mathbf{X}_c \stackrel{\text{def}}{=} \mathbf{U}_{1,2}$ the *common components*.

Finally we regressed $\mathbf{X}_c$ out from $\tilde{\mathbf{X}}_1^{\text{stim}}$ and $\tilde{\mathbf{X}}_2^{\text{stim}}$ separately, and used PCA (singular value decomposition if assuming zero mean) to extract the p_c -dimensional projections of the residual space respectively for Layer 1 and Layer 7. For example, we have the following for Layer 1

$$\mathbf{U}_{r1} \mathbf{D}_{r1} \mathbf{V}_{r1}^T \approx \left(\mathbf{I} - \mathbf{X}_c (\mathbf{X}_c^T \mathbf{X}_c)^{-1} \mathbf{X}_c^T \right) \tilde{\mathbf{X}}_1^{\text{stim}}$$

$362 \times p_c \quad p_c \times p_c \quad p_c \times p_c$

and we call $\mathbf{U}_{r1}$ *residual Layer 1*. Similarly, we obtained *residual Layer 7*

$$(\mathbf{U}_{r2}, \text{ where } \mathbf{U}_{r2} \mathbf{D}_{r2} \mathbf{V}_{r2}^T \approx \left(\mathbf{I} - \mathbf{X}_c (\mathbf{X}_c^T \mathbf{X}_c)^{-1} \mathbf{X}_c^T \right) \tilde{\mathbf{X}}_2^{\text{stim}}).$$

The cross-validation results can be found in Chapter 5.2.10 and 5.3.1 in Yang (2017). Given how *common components*, *residual Layer 1*, and *residual Layer 7* were created, we suggest the following intuitive interpretation. The *common components* are the linearly correlated components between the low-level Layer 1 features and the high-level Layer 7 features. Since the features in Layer 7 are highly relevant to the object-category labels, the *common components* represents the “object-category-relevant” or task-relevant low-level features (for example, informative edges that define the shape of an object). In contrast, *residual Layer 1* represents components in Layer 1 features that were roughly orthogonal to higher-level semantic information that is relevant to the task of object categorization. Finally, *residual Layer 7* represents linear components in Layer 7 that were roughly orthogonal to features in Layer 1; in other words, *residual Layer 7* represents unique high-level features beyond the information captured in the low-level features.

2.11 | Confidence intervals and statistical tests

2.11.1 | Percentile confidence intervals

After obtaining the statistics representing the regression effects (e.g., R-squared) for each time point, we computed the group-

averaged time series across participants, for which the confidence intervals were obtained through bootstrapping. We randomly resampled the time series of statistics at the participant level with replacement, and used the $(\alpha_0, 1 - \alpha_0/2)$ percentile confidence intervals of the bootstrapped samples (Wasserman, 2010). The significance level α_0 here was defined as $0.05/T/n_{\text{stat}}$, where T was the number of time points in the time series, and n_{stat} was the number of time series of statistics that were considered (i.e., Bonferroni correction). For example, in cases where we plot the R-squared for eight layers of features, $n_{\text{stat}} = 8$.

2.11.2 | Permutation-excursion tests

When examining whether a time series of statistics is significantly different from the null hypothesis in some time window, it is necessary to correct for multiple comparisons across different time points. Here, permutation-excursion tests (Maris & Oostenveld, 2007; Xu, Sudre, Wang, Weber, & Kass, 2011) were used to control the family-wise error rate and obtain a global p -value across time windows. In a one-sided test that examines whether some statistics were significantly larger than the null, we first identified clusters of continuous time points where the statistics were above a threshold, and then took the sum within each of these clusters. Similarly, in each permutation, the statistics of permuted data were thresholded, and summed within each of the detected clusters. The global p -value for a cluster in the original, non-permuted case was then defined as the proportion of permutations where the largest summed statistics among all of the detected clusters was greater than the summed statistics in the cluster from the non-permuted case.

To test whether the mean time series of some statistics (e.g., the R-squared in linear regression) across participants was significantly larger than that in the baseline time window before the stimulus onset, we computed the difference, separately for each participant, between the original time series of the statistic and the temporal mean of the statistics within the baseline time window. Then across participants at each time point, we used the t -statistics defined in Student's t -tests to examine if the group means of these differences were significantly above zero in any time window. That is, we used these t -statistics in the permutation-excursion test. Each permutation was implemented by assigning a random sign to the difference time series for each participant. This test, which we refer to as a *permutation-excursion t -test* hereafter, was implemented in **MNE-python**, where the number of permutations was set to 1,024. The threshold for the t -statistics in the permutation excursion test was equivalent to an uncorrected p -value ≤ 0.05 .

2.11.3 | Other corrections of multiple comparisons

In addition to the permutation-excursion tests discussed above, in cases where the possible dependence structure was not as readily characterized as compared to that in adjacent time points, we relied on a different method to correct for multiple comparisons—controlling the false discovery rate using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995).

2.12 | Data availability

The preprocessed data that support the findings in this study is openly available at Figshare (https://figshare.com/articles/MEG_scene_data/7991615). The python scripts that generated the results, as well as the link to the Figshare dataset and corresponding descriptions, are available at https://github.com/YingYang/MEG_Scene.

3 | RESULTS

By regressing neural activity at different time points and brain areas onto stimulus features extracted from a pretrained CNN, we obtained spatiotemporal correlation profiles that offer insights as to information flow in the human vision cortex. Below we present these correlation profiles in both the MEG sensor space and the source space, which was defined on the cortical surface. As an overview, we first examined how each layer of **AlexNet** accounted for the variance of neural activity. Next, we decomposed low-level and high-level **AlexNet** features into three orthogonal sets: the common components between low-level and high-level features, the residual low-level features and the residual high-level features; we present the neural correlation profiles with these sets and analyze how these profiles not only support feedforward information flow but also indicate non-feedforward information flow when participants process naturalistic scene images.

3.1 | Sensor-space regression

3.1.1 | Correlating the sensor-space neural activity with different layers in AlexNet

To examine whether the features extracted from different layers in the pretrained CNN (**AlexNet**) are able to explain the MEG data in the sensor space, we ran an ordinary least square regression analysis at each sensor and each time point for each participant. This approach allowed us to compare the linear dependence (or correlation) between different feature sets and the neural activity at different time points. For this regression, the neural responses to all 362 images were included.

Because **AlexNet** features are high-dimensional as compared to the number of observations, we need to avoid over fitting through either dimensionality reduction or other regularization methods. In that our main goal was to test whether a significant amount of variance was explained by each layer and for computational simplicity, we used PCA to reduce the dimensionality and included the first 10 principal components as regressors. The choice of 10 was admittedly arbitrary; we expect similar results so long as the number of components is neither too small (e.g., 2–3) nor too large such that the regression models overfit. We used the first 10 principal components of both the intact features and the *local-contrast-reduced features* from each layer as regressors; as mentioned in Materials and methods, in the *local-contrast-reduced features*, the variation capturing contrast in local neighborhoods at different locations of an image was removed.

We quantified the correlation between neural activity and the regressors in each layer as the proportion of variance explained (i.e., R-

squared). In other words, R-squared is the statistic in our correlation profiles. For purposes of visualizing overall effects, R-squared was averaged across all sensors at each time point for each participant. Figure 3 illustrates these results, where each curve represents the R-squared for each layer, averaged across all sensors and all participants. The transparent bands show the confidence intervals obtained by bootstrapping the observed R-squared time series at the participant level. *Permutation-excursion t-tests* were used to test whether the averaged R-squared across sensors was greater than the temporal average of that in the baseline time window (–140 to –40 ms with regard to the stimulus onset), during which MEG signals should be independent of the stimulus images, and thus the regressors. The significant time windows were identified where the *p*-values of the *permutation-excursion t-tests* were smaller than 0.05/8. These significant windows are marked by the colored lines under the curves in Figure 3. Note that these *p*-values were already corrected for multiple comparisons at different time points; the denominator eight was used as a correction for the eight tests corresponding to the eight layers according to the Bonferroni criterion.

In Figure 3, the left plot (Figure 3a) shows the results from using the intact features, while the right plot (Figure 3b) shows the results from using the *local-contrast-reduced features*. In both plots, we identified statistically significant time windows (from as early as about 60 ms to as late as about 600 ms), where the variance explained by the features was significantly greater than that in the baseline time window. These results indicate that the neural responses recorded by MEG in these time windows were correlated with the **AlexNet** features.

Because we used the same number of orthogonal principal components for each layer, the model complexity of the regression for each layer is the same. In this context, we can qualitatively compare the R-squared statistics between layers and analyze which layer explained the neural responses better. Both plots show a pattern of early-to-late, lower-level to higher-level shift, where lower-level layers, especially Layers 1 and 2, explain a larger proportion of variance before 150 ms, and higher-level layers, especially Layers 6 and 7, explain a larger proportion of variance from about 150 to at least 280 ms. This pattern is consistent with feedforward information flow as assumed in the hierarchical models of the visual cortex. Interestingly, Layer 8 generally explained lower variance as compared to the other seven layers; no significant time window was detected in Figure 3b. One possibility is that the object categories represented in Layer 8 did not align with the basic, neurally-discriminable categories that humans naturally use, or that the neural activity associated with object category labels is weaker than the neural activity associated with visual features.

Also of interest, although the proportion of variance explained by the principal components from the intact features appeared higher than that for the *local-contrast-reduced features*, the early-to-late, lower-level to higher-level temporal patterns were much clearer in the latter case. These results indicate that using the *local-contrast-reduced features* may facilitate better differentiation in the neural correlations with feature sets at different levels. As such, throughout the remainder of this article, we present only results with the *local-contrast-reduced features*; that is, unless there is a specific note, “features” will refer to the *local-contrast-reduced features*.

Next, we illustrate spatially which sensors demonstrated strong correlations with the **AlexNet** features by visualizing the individual traces of the variance explained in each sensor location across the topological maps. There were 102 unique sensor locations, each representing a one-magnetometer-two-gradiometer triplet. For visualization, we took the average of the proportion of variance explained for the three sensors in each triplet and plotted the averaged values at each time point. Note that the same sensor location could map to slightly different locations for different participants, due to individual variations in head sizes and head locations in the MEG helmet.

As a visualization, these results are purely qualitative and are intended to aid in a better understanding of the overall patterns of our results. We will present more rigorous statistical comparisons of the spatiotemporal correlation profiles in the source-space regression analysis later.

Figure 4 shows such a visualization—in each subplot, the lower panel depicts individual traces of the variance explained at each sensor location; the upper panel depicts the topological plots of the corresponding values at several time points. We show the proportion of variance explained by the first 10 principal components of the features in Layers 1, 3, 5, and 7; Layers 2, 4 and 6 demonstrated a similar pattern to their corresponding adjacent layers (not shown). The strongest correlation effects were in the posterior sensors, which are close to the visual cortex. From Layer 1 to Layer 7, we observe a shift of the correlation effects from early windows to late windows, especially from 50 to 250 ms. For Layer 1 (and also Layer 3), it is interesting that after the first transient peak near 80 ms, there appears to be a second set of peaks from about 280 to 380 ms. Because each stimulus image was presented for 200 ms, we posit that this second set of peaks may arise as a neural response to the disappearance of each image at 200 ms—there was roughly an 80 ms delay from the stimulus onset to the first peak, so we might expect a similar delay in the response to the disappearance of images.

3.1.2 | Correlating the sensor-space neural activity with three sets of decomposed features derived from Layer 1 and Layer 7

Using Layer 1 as low-level features and Layer 7 as high-level features, we further derived three orthogonal sets of features from these starting points, using the CCA of the *local-contrast-reduced features* ($p_c = 6$, see Materials and methods): the *common components* represents the shared linear components between Layer 1 and Layer 7, or the components in the low-level features that are highly correlated with the high-level features; the *residual Layer 1* represents low-level features that are roughly orthogonal to Layer 7; the *residual Layer 7* represents high-level features that are roughly orthogonal to Layer 1. In Figure 5, for each of the three sets of derived features, we present the averaged traces of the proportion of variance explained for the triplet at each sensor location and the topological maps at several time points. Again, it is worth noting that this is a qualitative visualization. We present the statistical comparisons between different feature sets in the source-space ROI analysis.

As illustrated in Figure 5, the *residual Layer 1* exhibited an early peak spanning from 60 to 120 ms and centered near 80 ms (Figure 5a), while the *residual Layer 7* exhibited correlation effects later, from about 120 to 380 ms (Figure 5c). For an easy comparison, the scale for the topological maps and the traces were set to the same for the *residual Layer 1*, the *residual Layer 7*, and the *common components*; the temporal changes in the topological visualization for *residual Layer 7* was not as obvious as for the *residual Layer 1* but was still noticeable. The *common components* appeared to have two sets of peaks, one spanning from 60 to 130 ms and centered near 80 ms, and the other spanning from about 280 to 380 ms with the local maxima near 300 and 350 ms (Figure 5b). This latter set of peaks may correspond to neural responses to the disappearance of each stimulus at 200 ms. Although some posterior sensors also showed a tiny late peak for *residual Layer 1* near 270 to 330 ms (see the colored traces in Figure 5a), this peak did not last till 350 ms as for the *common components* in Figure 5b). These results indicate that near 350 ms the *common components* (the low-level components that are correlated with high-level features) could be represented differently from the *residual Layer 1* (the low-level components that are roughly orthogonal to high-level features).

In Figure 5d, for comparison, we also show the correlation effects for the first six principal components of the *local contrast features* (with a different scale than the three upper plots to illustrate the temporal trends). Note that the local contrast features were orthogonal to the three sets of features above (see Materials and methods). The local contrast features showed large correlation effects spanning from 50 to 400 ms, with an early peak centered around 100 ms and a set of later peaks within 280 to 380 ms. These correlation effects were larger than those for the three sets above, indicating that local contrast did elicit relatively strong neural responses.

Summarizing the regression results in the sensor space, we observed an early-to-late, lower-level to higher-level shift in the temporal correlation patterns, indicating a feedforward information flow during visual scene perception. When we decomposed the *local-contrast-reduced features* from Layers 1 and 7 into three roughly orthogonal feature sets—the *common components*, the *residual Layer 1* and the *residual Layer 7*—we observed an apparent temporal separation of the low-level and high-level features (e.g., when comparing the *residual Layer 1* and the *residual Layer 7* or comparing the *common components* and the *residual Layer 7*). We also observed some later correlation effects starting near 280 ms, which had different dynamics for the *residual Layer 1* and the *common components*. In addition, the *local contrast features* explained a larger proportion of variance in the neural data than the three feature sets derived from the *local-contrast-reduced features*, suggesting that failing to partial out local contrast features from **AlexNet** features would result in majority of observable correlation effects being due to the local contrast features.

3.2 | Source-space regression

As presented above, the regression results in the sensor space are only informative regarding temporal correlation profiles between the neural activity and the stimulus features. Here, we shift to the source

FIGURE 4 The proportion of variance (%) explained by the features from Layers 1, 3, 5, and 7. In each sub-figure, the curves plotted in the lower panel show the proportion of variance explained for each time point and each sensor location. The colors of the curves correspond to the color-coded sensors in the upper-left corner of the plot. The upper panel shows topological maps of the values at the marked time points. Note that the values for the three sensors in a one-magnetometer-two-gradiometer triplet at each location were averaged in these plots [Color figure can be viewed at wileyonlinelibrary.com]

space to examine the spatiotemporal correlation profiles. Source localization was achieved using the dSPM (Dale et al., 2000) method; in this way, neural activity was obtained at each source point in the brain space as defined on the cortical surfaces. We then ran ordinary least

square regression analyses to obtain the correlation profiles. Note that source localization per se is a challenging problem, as the number of source points in the brain space (about 8,000 in our case) is an order of magnitude larger than the number of sensors (306). The

FIGURE 5 The proportion of variance (%) explained by the the common components, the residual Layer 1, residual Layer 7, and the local contrast features. In each subfigure, the traces plotted in the lower panel show the proportion of variance explained for each time point and each sensor location. The colors of the curves correspond to the color-coded sensors in the upper-left corner of the plot. The upper panel shows topological maps of the values at the marked time points. Note that the values for the three sensors in a one-magnetometer-two-gradiometer triplet at each location were averaged in these plots [Color figure can be viewed at wileyonlinelibrary.com]

dSPM method uses a type of regularization to obtain a unique solution, which is equivalent to placing a zero-mean, independent and identically distributed Gaussian prior on the source neural activity

given that we have no further information about where to localize the signals. To verify that our results are not due to this specific source localization prior, in Appendix, we also applied an L_1/L_2 -based

FIGURE 6 The proportion of variance explained by different layers, computed from the dynamic statistical parametric mapping (dSPM) source solutions that were morphed onto a common template and averaged across participants [Color figure can be viewed at wileyonlinelibrary.com]

sparsity-inducing source regression method—a short-time Fourier transform regression model (STFT-R, Yang et al., 2014)—and obtained qualitatively similar results.

3.2.1 | Whole-brain visualization of the source-space regression

Subsequent to source localization and regression, we mapped the source space of each participant onto a default template in **FreeSurfer**. We then obtained whole-brain maps of the averaged R-squared statistics for the first 10 principal components for each layer in **AlexNet** (the same as those used in the “sensor-space regression”). Figure 6 shows the visualization at different time points from a ventral view of the cortical surfaces for Layers 1, 3, 5, and 7.

Adjacent source points typically contribute to the sensor recordings in a similar way; as a result, the regularization in dSPM generally have a spatial blurring effect, where an underlying single large current dipole can be reconstructed as distributed current dipoles covering a large cortical area. As a consequence, when we observe strong effects (i.e., large averaged R-squared values) in large areas, this pattern may be due to either local large effects or genuinely distributed effects. Keeping this issue in mind, we examine our results in Figure 6. At 60 ms, the correlation effects localized in the posterior end near the EVC were strongest for Layer 1, and for Layers 3, 5, and 7, the correlation effects gradually decreased; at 100 ms, the regression effects had large magnitudes and spread over large areas of the ventral visual cortex for all layers, although the effects were stronger in Layers 1 and 3 than in Layers 5 and 7; at 140 to 280 ms, the correlation effects appeared stronger and in larger areas for Layers 3, 5, and 7 than for Layer 1, and the effects spread to more anterior regions for Layers 5 and 7. This overall pattern is consistent with feedforward

information flow along the hierarchy from posterior to anterior regions of the visual cortex. In addition, at ~300 to 380 ms, we observed a second correlation peak for most layers in posterior regions close to EVC. As mentioned above, this later peak may be a result of a correlation between the neural responses to the disappearance of the images and the CNN features of those images.

3.2.2 | Statistical analysis of the source-space regression results in ROIs

Next, we present the results of the source-space regression analysis using the three sets of derived features, the *common components*, the *residual Layer 1* and the *residual Layer 7*. After obtaining the dSPM source estimates, we ran a linear regression for each source point at each time point, using each of the three feature sets as the regressors to obtain the R-squared statistics. Each time series of the R-squared values may potentially have different magnitudes across different source points and participants, so we further normalized these time series using the following method. For each source point, we divided the R-squared values at each time point by the sum across all time points and all three feature sets.

In order to conduct rigorous statistical tests to compare the correlation profiles between these three feature sets, we focused on several representative ROIs along the visual hierarchy, including the pericalcarine areas that covered the early visual cortex (EVC) and the object/scene-selective areas (LOC, PPA, RSC, and TOS). This ROI approach was in contrast to a whole-brain analysis, where we would have needed to control for multiple comparisons at thousands of source points. In particular, by aggregating the correlation effects within each ROI, we had fewer multiple comparisons to correct for and therefore higher statistical power. Since these regions were

FIGURE 7 Correlation profiles (the normalized proportion of variance explained by the linear regression) in the regions of interest (ROIs) in the source space. First column: The averaged normalized statistics of the correlation effects across participants (color code: blue: *residual Layer 1*; green: *residual Layer 7*; red: *common components*). Second to fourth columns, pairwise differences between the three feature groups (color code: Cyan: *residual Layer 7-residual Layer 1*; yellow: *residual Layer 1-common components*; magenta, *residual Layer 7-common components*). The transparent bands indicate bootstrapped confidence intervals at the participant level. Gray areas indicate time windows where the differences were significantly nonzero, after correction for multiple comparisons for all time points from 0 to 900 ms and the number of ROIs [Color figure can be viewed at wileyonlinelibrary.com]

defined individually for each participant, we also did not have to morph the individual source spaces onto a template. Our analysis proceeded by averaging these normalized R-squared values across the source points within each ROI. Note that these normalized values reflect the relative correlation effects between localized neural activity and the three groups of regressors within the ROI, and thus they represent the spatiotemporal profiles of the correlation effects we will focus on hereafter.

Figure 7 shows the correlation profiles and related comparisons within each ROI. The first column shows the average of the normalized R-squared values across participants for each group of regressors within each ROI. The transparent bands show 95% confidence

intervals, bootstrapped at the participant level and corrected for multiple comparisons across all time points and for the three regressor groups using the Bonferroni criterion. We also ran pairwise comparisons between the correlation profiles of the three feature sets. In these pairwise comparisons, for each source point, each time point, and each participant, we computed the ratio of the R-squared value for each regressor set (i.e., each feature set) to the sum of the R-squared values across the three regressor groups. In this way, each time point had comparable statistics that described the relative strength of the correlation effects for each of the three regressor groups. Again, for each participant, we averaged these ratios across the source points within each ROI. We then took pairwise differences

(*residual Layer 7-residual Layer 1*, *residual Layer 1-common components*, and *residual Layer 7-common components*), and examined whether the averaged differences across participants were significantly different from zero in each ROI. The remaining three columns in Figure 7 show the averaged differences between each pair of the three sets of features: the second column (cyan) corresponds to the difference between *residual Layer 7* and *residual Layer 1*; the third column (yellow) corresponds to the difference between *residual Layer 1* and *common components*; the fourth column (magenta) corresponds to the difference between *residual Layer 7* and *common components*. The transparent bands show the bootstrapped 95% confidence intervals, corrected for the three comparisons at all the time points using the Bonferroni criterion. *Permutation-excursion t-tests* were used to identify time windows where the two-sided *p*-values were smaller than .05. Note that in this case, we only corrected for multiple comparisons across time points. To further correct for comparisons across different pairs and multiple ROIs, we controlled the false discovery rate (FDR) at 0.05 using the Benjamini-Hochberg procedure. The gray boxes in Figure 7 indicate the time windows that survived the correction (i.e., the difference between the correlation effects for the two sets of features was significantly greater or smaller than zero), and the *p*-values of the *permutation-excursion t-tests* (before the FDR correction) are marked. Note that there appeared to be some time windows that did not survive the correction, but in which we could visually observe some possibly nonzero differences based on the confidence intervals (which were not corrected for multiple ROIs). However, we were not able to claim that these windows had significantly nonzero differences in the comparisons. Nevertheless, even if the differences at some time points were not significant, there may still exist true underlying differences that we were unable to detect given the statistical power of our study.

Next, we analyze the correlation profiles and the pairwise comparisons for each ROI. In addition to contrasting the patterns between the three groups within each region, we also focus on the temporal changes of the correlation effects. We can also qualitatively compare correlation profiles in the EVC and in the object/scene selective regions (e.g., LOC, PPA).

In the EVC (in the first row and the first column of Figure 7, or see Figure 8 for an enlarged view), the *residual Layer 1* and the *common components* (the blue and red curves respectively)—corresponding to the low-level features roughly orthogonal to high-level features, and the low-level features that were correlated with high-level features (i.e., object-category-relevant features)—had early transient correlation effects within 60 to 120 ms, peaking near 80 ms. The correlation effects of the *residual Layer 7* (the green curve)—corresponding to the high-level features that were roughly orthogonal to low-level features—increased later, starting at about 100 ms, plateaued near 140 ms to about 300 ms, and gradually decreased near 300 to 400 ms.

From the pairwise comparison between the *residual Layer 7* and the *residual Layer 1* (the cyan curve in the second column and the first row of Figure 7), we can see a negative peak within 60 to 100 ms, indicating that the correlation effect of the *residual Layer 7* was smaller than that of the *residual Layer 1*. Following this time window,

FIGURE 8 Correlation profiles in the early visual cortex (EVC) in the source space (enlarged from Figure 7). [Color figure can be viewed at wileyonlinelibrary.com]

the difference became positive from about 120 to 260 ms, indicating that the correlation effect of the *residual Layer 7* was greater than that of the *residual Layer 1* in this window. These results are consistent with the patterns of the blue and green curves shown in the first column, which reveal an early-to-late shift of the correlation effects from low-level to high-level features.

Notably, in the EVC, the *residual Layer 1* and the *common components* both exhibited a second peak starting near 280 ms (in addition to the first peak at 60 to 120 ms, see the blue and red curves in Figure 8). However, this second peak appeared smaller and more transient for the *residual Layer 1* than for the *common components*; the correlation effect also lasted longer (until at least 380 ms) for the *common components* (the red curve) than for the *residual Layer 1* (the blue curve). This can be verified with the significant negative window in the pairwise comparison between the *residual Layer 1* and the *common components* near 300 to 380 ms (see the yellow plot in the first row and the third column of Figure 7). In contrast, around the first transient peaks (60 to 180 ms) for the *residual Layer 1* and the *common components*, the difference (*residual Layer 1-common components*) was positive at the beginning and negative later, indicating that the correlation effect with the *residual Layer 1* was higher, but that the correlation effect with the *common components* lasted slightly longer than the *common components* in the falling phase of the blue and red peaks as shown in the first column of Figure 7.

In the pairwise comparison between the *residual Layer 7* and the *common components* (the magenta plot in the first row and the the fourth column of Figure 7), we observed negative peaks from about 60 to 120 ms and near 330 to 380 ms, corresponding to the two transient red peaks of the *common components* in the first column. There are some significant positive differences near 260 ms, indicating higher correlation effects with the high-level features (*residual Layer 7*) as compared to the object-category-relevant low-level features (the *common components*).

One of the key findings of this study is the different correlation effects observed in the EVC with respect to the *residual Layer 1* and

the *common components* from 300 to 380 ms. This result indicates that the EVC is able to differentiate object-category-relevant and less-object-category-relevant low-level features. The specific timing of these second peaks near 280 ms may be due to the disappearance of the stimuli at 200 ms—the EVC was sensitive to changes of visual inputs that were related to the low-level image features that disappeared. Nevertheless, compared with the first peaks of the blue and red curves near 60 to 120 ms (in Figure 8), which were due to the responses to the stimulus onset, the dynamics of these second peaks was different. In particular, in the early peaks, the correlation effect with the *residual Layer 1* (blue) was larger and the correlation effect with the *common components* (red) was just slightly longer than the blue peak; whereas in this later time window, the red peak was higher and lasted much longer. If the late red and blue peaks starting at 280 ms are due to the disappearance of the stimuli, some non-feedforward dynamics, either due to top-down feedback or local lateral connections, appears to have “guided” the EVC to focus more strongly and for longer on the object-category-relevant low-level features.

In an object-selective region, the LOC (in the second row of Figure 7), which is at a higher level than the EVC along the hierarchy, the scales of the early transient peaks (60 to 130 ms) for the low-level features (the *residual Layer 1* and the *common components*, the blue and red curves) were comparable to the scale of the later wide peaks (150 to 400 ms) for the high-level features (*residual Layer 7*, green), which was in contrast with the pattern in the EVC. There appeared to be a second red peak near 350 ms for the *common components*; however, it was not much larger than the effect for the *residual Layer 7* (green) in the same time window as observed in the EVC.

In the LOC, the pairwise difference plotted in the second column of Figure 7, (the cyan plot in the second row, *residual Layer 7-residual Layer 1*) shows a similar pattern to that in the EVC, but the early negative difference was smaller and the later positive difference lasted longer, corresponding to the profiles of the correlation effects in the first column of Figure 7. In the pairwise difference plotted in the third column of Figure 7, (the yellow curve in the third row, *residual Layer 1-common components*), there was both an early and a late negative windows, indicating that the correlation effect of the *common components* was larger than that of the *residual Layer 1*. These results are consistent with the hierarchical organization for the visual cortex, in that the LOC, being at a higher level in the hierarchy, showed higher correlation effects with the object-category-relevant low-level features (the *common components*) than with the low-level features that were roughly orthogonal to high-level features (the *residual Layer 1*). In the pairwise difference plotted in the fourth column of Figure 7, (the magenta plot in the fourth row, *residual Layer 7-common components*), we observed an early negative window and a later positive window, indicating a temporal shift from a higher correlation effect with the object-category-relevant low-level features to a higher correlation effect with the high-level features, again consistent with feedforward information flow.

One might have expected that the LOC, which is at a higher level within the hierarchy than the EVC, would show smaller correlation

effects with low-level features than what we observed. We speculate that because the columns in the forward matrix in source localization inherently have strong spatial correlations, the reconstructed source solutions can be spatially blurred, and the correlation effects with low-level features can “leak” from lower-level visual areas into the LOC. Even in the absence of spatial blurring, the hierarchy can be gradual such that we may only observe small relative differences between the correlation profiles for the EVC and the LOC.

The third row Figure 7 shows the results for a scene-selective ROI, the parahippocampal place area or PPA. In the first column of Figure 7, the correlation profiles look similar to those in the LOC. In the pairwise comparison in the second column (*residual Layer 7-residual Layer 1*, cyan), we can also observe an initial negative window and a short later positive window, indicating an early higher correlation with the *residual Layer 1* and a later higher correlation with *residual Layer 7*. In the pairwise comparison in the third column (*residual Layer 1-common components*, yellow), we did not observe any significant differences. In the pairwise comparison in the fourth column of Figure 7 (*residual Layer 7-common components*, magenta), we again observed an early negative window and a later positive window, indicating early higher correlation with the object-category-relevant low-level features as compared to the high-level features, and a reversed pattern at later time points. Interestingly, this detected positive window is near 500 ms, perhaps indicating sustained processing of high-level features in the PPA.

The fourth row of Figure 7 shows the results in a scene-selective ROI near the TOS. In the first column, the correlation profiles appear to be in the middle of a transition from the pattern in the EVC to the patterns of the LOC and the PPA. This may be because the TOS was close to the EVC in euclidean distance in the source space. In the pairwise comparison in the second column of Figure 7 (*residual Layer 7-residual Layer 1*, cyan), again, we see an initial negative window and later positive windows, indicating an early higher correlation with the *residual Layer 1* and a later higher correlation with the *residual Layer 7*. The latest significant positive window is at 500 ms, which may indicate some sustained processing of high-level features. In the pairwise comparison in the third column (*residual Layer 1-common components*, yellow), we observed a positive but not significant window from about 60 to 100 ms, and several negative windows within the range 120 to 320 ms, which indicates a shift of the correlation effect from the less-object-category-relevant low-level features to the object-category-relevant low-level features, consistent with feedforward information flow. In the pairwise comparison in the fourth column (*residual Layer 7-common components*, magenta), we observe an early negative window and a later positive window, indicating an early higher correlation with the object-category-relevant low-level features as compared to the high-level features and a later higher correlation with the high-level features.

The fifth row of Figure 7 shows the results in a scene-selective ROI near the retrosplenial complex (RSC). In the first column, the correlation profiles appear similar to those in the LOC and the PPA. In the pairwise comparison in the second column of Figure 7 (*residual Layer 7-residual Layer 1*, cyan), we see an initial negative window near

60 to 100 ms and a later positive window near 130 to 170 ms. This pattern is consistent with an early-to-late shift from the less-object-category-relevant low-level features to high-level features. In the pairwise comparison in the third column (*residual Layer 1-common components*, yellow), we observe a positive window near 70 to 90 ms, and a negative window near 110 to 140 ms. This pattern is consistent with an early-to-late shift from the less-object-category-relevant low-level features to object-category-relevant low-level features. In the pairwise comparison in the fourth column (*residual Layer 7-common components*, magenta), we observed a negative window near 70 to 120 ms, which indicates a higher correlation effect with object-category-relevant low-level features than with the high-level features in this early time window. In general, the results of these pairwise comparisons are consistent with feedforward information flow.

3.2.3 | Classifying semantic categories of the images using the low-level features and neural activity in the EVC

In the whole-brain and ROI-based source-space analysis above, we have observed differential correlation profiles with the three sets of features—the *common components*, the *residual Layer 1* and the *residual Layer 7*. Interestingly, in the EVC, we observed higher neural correlation with the *common components* than with the *residual Layer 1* around 300 to 380 ms after the stimulus onset. Next we ask the question “can such differences in the correlation profile be related to how human classify the images semantically?”

To answer this question, we ran a simple decoding analysis to evaluate whether the three sets of six dimensional features were able to predict some semantic category labels (e.g., “indoor” vs. “outdoor”) of the stimulus images. We manually created three semantic labels for each of the 362 images: “is it an indoor scene?,” “is it a natural outdoor scene?” and “is it a man-made outdoor scene?”. For each question, there were two classes, 0 if “no” and 1 if “yes.” For each image, there is only one “1” for the three questions. These three questions, which were used in organizing the SUN database (Xiao, Hays, Ehinger,

Oliva, & Torralba, 2010), provided a high-level semantic categorization of the scene images. For each of the labels (“indoor,” “natural outdoor,” and “man-made outdoor”), we used the six-dimensional features—the *common components*, the *residual Layer 1* and the *residual Layer 7*, as well as the first six principal components of the “local contrast features”—to classify the labels with a linear support vector machine (with a fixed penalization parameter $C = 10.0$). A 10-fold cross validation was used and the receiver operating characteristic (ROC) curve of the testing results in each fold was computed and averaged.

Figure 9 shows the ROC curves (blue for the *residual Layer 1*, red for *common components*, green for *residual Layer 7* and black for “local contrast features”). The area under the ROC curves (AUC), which indicates how much information the feature set contains about the class labels, was also displayed in the legends—AUC = 0.5 at the chance level and AUC = 1.0 at perfect classification. For all the three labels, the ROC curves of the *common components* (red) were above the chance line and above the ROC curves for the *residual Layer 1* in blue; the AUC of the *common components* (0.61 to 0.79) was higher than that of the *residual Layer 1* (≤ 0.5). Although both of these two sets are derived from Layer 1, the *common components* included components that were correlated with Layer 7 features, and thus they were likely to include more relevant information to object labels and thus semantic information about the images. In comparison, the *residual Layer 1* was orthogonal to *common components* and was less likely to contain as much semantic information. The results in the decoding analysis supported the hypothesis above.

The *residual Layer 7* (the green curves), which was derived from Layer 7, generally had an AUC >0.5 , but was not always higher than the *common components*. Interestingly, the ROC curves of the “local contrast” features, which we regressed out before deriving the three sets of features, were not as high as the common features—they were not as useful in terms of classifying these semantic labels.

The EVC demonstrated differential neural correlation with the two types of low-level features, the *residual Layer 1* and the *common components*, at 300 to 380 ms. Does the neural activity at this

FIGURE 9 The receiver operating characteristic (ROC) curves and the area under the curve (AUC) for classifying semantic labels using the three feature sets and neural activity in the early visual cortex (EVC). (red: common; blue: residual Layer 1; green: residual Layer 7; black: the local contrast features; yellow: the EVC activity averaged across 300–380 ms) [Color figure can be viewed at wileyonlinelibrary.com]

window also encode semantic information about the images? We ran the decoding analysis on the source-localized neural activity in the EVC averaged across 300 to 380 ms for each participant, with the same type of classifiers and the same cross-validation paradigm as above. The number of source points in the EVC varied from 46 to 88 across participants; in the support vector machine classifiers, the penalization parameter was still set to 10.0. The average ROC curves and the bootstrapped confidence intervals (95%, after correction for 100 points on each ROC curve) are shown in yellow in Figure 9, which was above the chance line for “indoor” and “natural outdoor” and marginally above chance for “man-made outdoor.” These results indicate that the source-localized MEG responses in the EVC in this time window, which was correlated with the *common components*, also encoded some semantic information and could be something useful for scene classification.

To summarize these results, using a regression analyses in the source space, we obtained spatiotemporal correlation profiles between the neural activity and the three groups of features (the *common components*, the *residual Layer 1* and the *residual Layer 7*). By analyzing these profiles, we observed progressive shifts from early (60 to 120 ms) to later (roughly after 120 ms) time windows, from lower-level to higher-level features, and from low-level regions to higher-level regions along the hierarchy. These results strongly support a model of visual cortex in which feedforward information flow is intrinsic to perceptual processing. Perhaps more *novel* is our observation that at a later time window from about 300 to 380 ms the correlation of the EVC with the object-category-relevant low-level features (the *common components*) was larger than the correlation with the less-object-category-relevant low-level features (the *residual Layer 1*). This result suggests that a non-feedforward process (e.g., top-down influences) may facilitate EVC in distinguishing between the low-level features carrying different kinds of information and in representing those low-level features that are object-category-relevant.

4 | DISCUSSION

One of the advantages of MEG (as compared to EEG or fMRI) is that it allows for measurement of joint spatiotemporal patterns of neural activity. Therefore, we are somewhat surprised that the majority of previous work using MEG to correlate neural responses with computer vision features focused mainly on temporal patterns (Cichy, Khosla, Pantazis, & Oliva, 2017; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016c; Clarke et al., 2014). In line with the standard hierarchical model of primate visual cortex, these prior studies observed an early-to-late shift from lower-level to higher-level features—a result consistent with the results we present here. Additionally, when joint spatiotemporal patterns were considered, Clarke et al. (2014) observed correlations between neural activity and visual/semantic features in source space indicating feedforward information flow—again, consistent with our present results. Using a somewhat different approach, Cichy, Pantazis, and Oliva (2016) “fused” fMRI and MEG recordings by comparing the representational similarity of visual objects in the

two imaging modalities with no reference to externally-derived features; they too observed an early-to-late shift in a feedforward direction of the cortical hierarchy.

What our study adds to this literature are more comprehensive spatiotemporal profiles of visual cortical processing; in particular, we rely on features derived from a more sophisticated computer vision model (**AlexNet**) as compared to the model used in (Clarke et al., 2014) (for a comparison of such models, see Yamins et al., 2014). Moreover, we included a relatively large number of complex, naturalistic scene images (rather than single objects on blank backgrounds as used in Clarke et al. (2014) and Cichy, Pantazis, & Oliva, 2016).

Finally, and perhaps most uniquely, we developed a novel decomposition of low-level and high-level features, and from them derived three orthogonal sets of features: (a) the common components between the low-level and high-level features, which can be interpreted as “object-category-relevant” low-level features; (b) the residual low-level features that is roughly orthogonal to high-level features, which can be interpreted as “less-object-category-relevant” low-level features; and (c) the residual high-level features that are relevant to object categories and from which low-level components are partialled out. By comparing the spatiotemporal neural correlation profiles with these three sets, we provided new evidence for non-feedforward processing within human visual cortex.

The key observation concerns the temporal correlation profiles of the EVC with the two sets of low-level features: (a) the “object-category-relevant” low-level features and (b) the “less-object-category-relevant” low-level features. In the early neural responses to the stimulus images near 60 to 120 ms, the correlation effects with these two low-level feature sets had similar dynamics, peaking around 80 ms; the correlation effect with the “less-object-category-relevant” low-level features was stronger and decayed slightly faster than the “object-category-relevant” low-level features. In a later time window starting near 280 ms, which was likely to be the window of neural responses to the disappearance of the stimulus images at 200 ms, the neural correlation profiles were different between the two sets of features—the “object-category-relevant” low-level features had stronger and longer correlation effects with the neural activity in the EVC which lasted till about 380 ms, compared to the “less-object-category-relevant” low-level features, which only showed transient correlation effects. If the visual system only has pure feedforward processing, then the EVC is unlikely to differentiate between the two sets of features as much as seen in our results. A more likely explanation is that there is feedback information flow in the visual cortex, potentially from higher-level regions to the EVC that “emphasizes” the processing of the “object-category-relevant” information, or “guides” the EVC to separate “object-category-relevant” low-level features from other low-level features. Alternatively, such separation may result from self-organizing behavior (e.g., similar to the self-organizing map, Kohonen, 1990) via lateral connections locally in the EVC, which is essentially similar to a local feedback mechanism. In sum, this novel observation provides unique insights on the information flow during the processing of naturalistic images of scenes.

4.1 | Potential pitfalls and future directions

Based on our observation of a linear dependence between the low- and high-level features, we developed a novel method for separating features into three orthogonal sets: the “object-category-relevant” low-level features, the less-object-category-relevant low-level features, and object-category-relevant high-level features. In tandem with the high temporal resolution and reasonable spatial resolution afforded by MEG, we were able to observe evidence for both bottom-up (feedforward) and potentially top-down (feedback) mechanisms during naturalistic scene perception. However, as with all studies, there are specific limitations in our choice of experimental methods and analyses—next we discuss these challenges as well as relevant future directions.

4.2 | Neural responses to the disappearance of stimuli

In our experimental design, each stimulus image was presented for 200 ms and then disappeared—the screen switched back to a fixation cross (“+”) displayed against a gray background. This disappearance necessarily produced changes in the visual input, which will drive neural responses in the EVC. Liang, Shen, Sun, and Shou (2008) characterized the magnitudes of responses in cat V1 to the disappearance of stimuli, which were, somewhat unexpectedly, comparable to the magnitudes of responses to the appearance of stimuli. In our analyses, although the mean responses to the appearance and disappearance of all stimuli were subtracted from the data during preprocessing, the image-specific responses (reflected in deviations from the mean) may be correlated with features of each image. This correlation is one possible explanation for why we observed peaks starting near 280 ms in the effects associated with the *common components* and the *residual Layer 1* in the EVC (see Figure 8). Interestingly, from the perspective of the complete visual processing stream, the disappearance of stimuli should not be treated as equivalent to the appearance of stimuli, because in contrast to appearance, with disappearance there is no additional semantic information presented across the changes in the visual input. Consistent with this logic, the *residual Layer 7* did not show any increase in correlation effects from 300 to 400 ms—the time window where responses to stimulus disappearance are likely to occur.

Regardless of the potential cause of the later peaks in the correlation effects, the difference between the effects for the two sets of low-level features is worth discussing. In particular, the EVC exhibited a stronger and longer correlation effect with the *common components* (the object-category-relevant low-level features) as compared to the *residual Layer 1* (the less-object-category-relevant low-level features). If we assumed only feedforward processing, we would expect similar correlation effects for both groups of low-level features (as seen in the early correlation effects near 80 ms in Figure 8). Given that we observed a distinction in the EVC between the two feature sets, we posit that non-feedforward processes, for example, feedback from the

higher-level cortex to the EVC or lateral recurrent interactions within the EVC neurons, provide a better explanation of our results.

Note that disappearance of stimuli may evoke an aftereffect that is perceivable to participants. Hence, our results provide some insight for further tests of the above speculation via behavioral experiments. We hypothesize that if participants perceive a stimulus aftereffect due to top-down information flow (as opposed to, for example, a retinal aftereffect), then this aftereffect should be mainly driven by the object-category-relevant low-level features. As such, one could design stimuli to manipulate presence/absence of such features, thereby manipulating the degree of the aftereffect.

As a related point, our design used a relatively short presentation duration (200 ms) in order to reduce saccade artifacts; additionally, we did not mask our stimuli post presentation (e.g., using white noise patterns). The observed later peaks of correlation effects mentioned above may be considered as a result of these design limitations, where the disappearance of the stimulus images interfered with the intact dynamics of visual processing. However, the positive here is that our study with the current design also provides novel observations regarding feedback within the visual system—the specifics of which have not been discussed much in the literature as far as we are aware. Future experiments should examine how the differential coding of different types of low-level features (e.g., *common components* and *residual Layer 1*) change under varying stimulus durations and masking conditions.

4.3 | Local contrast

In the canonical correlation analysis (CCA) of the intact features from **AlexNet**, we observed that CCA components between Layers 1 and 7 have high correlations with the local contrast features (not documented in the main text but see Chapter 5 in Yang, 2017). Interestingly, in our initial preliminary analysis of the data, we did not partial out the local contrast features when obtaining the three groups (the *residual Layer 1*, the *common components* and the *residual Layer 7*). Under these conditions, the neural correlation effects for the *common components* were much higher than those for either the *residual Layer 1* or the *residual Layer 7*, presumably due to a large correlation of the *common components* with the local contrast features. Please see Supplementary Figure S7 for a visualization of the difference before and after removing the local contrast features. After removing the local contrast features, the correlation profiles (proportion of variance explained) became roughly comparable between the three feature groups in terms of magnitude.

In this sense, local contrast appears to be a confounding factor. Nevertheless, local contrast may be inherently included in the statistical regularities of naturalistic images. Intuitively, local patches with high contrast often contain informative features with respect to shape and boundaries—consequently they can be related to semantic information, although using only the first six principal components of the local contrast, we were unable to obtain decoding performance commensurate with that obtained using the common components in classifying semantic labels of the images. Future analysis of local contrast features, which

can be implemented with a large set of images, may help us better understand the statistical properties of naturalistic images. Moreover, in future studies, to alleviate the influence of this confounding factor, we can design new stimuli by adding high contrast features that are less relevant to semantic information, for example, irregular shadow contours. It would be interesting to study whether such high-contrast features are coded differently from the genuinely informative high-contrast features (e.g., true physical edges or contours) within the visual cortex, as well as in a CNN trained on “standard” naturalistic images. Moreover, one might also use a generative neural network (e.g., Goodfellow et al., 2014) to create experimental test stimuli that are similar in local contrast and other low-level features to naturalistic images but that do not contain identifiable objects.

4.4 | Regression analysis without data split

In the Results section, we measured the R-squared or the proportion of variance explained without doing a data split. This was in a framework of statistical hypothesis testing; we viewed the proportion of variance explained by linear combinations of the 6 or 10 dimensional features as a testing statistic, and compared these statistics at different time points (e.g., comparing time points after and before the stimulus onset) and between different feature sets. The null hypotheses was that the proportion of variance explained during 0–900 ms was the same as baseline before the stimulus onset, and that the proportion of variance explained by different feature sets were the same. We observed significant differences in testing these null hypotheses in the Results section.

Nevertheless, it is worth noting that this proportion of variance explained is a biased estimator in a non-data-split context, and the bias increases with the number of feature dimensions (i.e., the number of regressors). The larger the feature dimension is, the linear model overfits more and the proportion of variance explained gets larger. However, when the number of features is small, this bias issue is less of a concern; moreover, we can make the comparison fair by using a constant number of feature dimensions while keeping the different dimensions of features orthogonal to each other. In our analysis, the feature dimension (6 or 10) was much smaller than the number of observations (362); additionally, we always compared the proportion of explained variance by the same number of orthogonal features (10-dimensional features in each of the **Alexnet** layers and 6-dimensional features in the *common components*, the *residual Layer 1* and the *residual Layer 7*). Therefore, we think our analyses are valid even without doing data split.

With that being said, we did run a cross-validation analysis to obtain a theoretically unbiased estimate of the proportion of variance explained by the features. An 18-fold cross validation was used—in each fold, about 20 of the 362 observations (corresponding to the 362 images) were held out for testing and the remaining were used for training (i.e., fitting the linear models). Suppose the neural data corresponding to the testing observation at a time point was $\mathbf{Y}$, and the predicted response was $\hat{\mathbf{Y}}$, then we defined the relative error (RE) as $(\|\hat{\mathbf{Y}} - \mathbf{Y}\|_2^2 / \|\mathbf{Y}\|_2^2)$, and defined $(1 - \text{RE}) \times 100\%$ as the cross-

validated version of proportion of variance explained. Note that this value can be negative, because there may be a discrepancy in the training and testing data—after fitting the regression, there could be a nonzero bias in $\mathbf{Y}$, which can make the RE greater than 100%, especially when the training and testing sets are small as in our case.

Supplementary Figure S6 shows the cross-validated version of the explained variance (in the right column) in the ROIs and the non-data-split explained variance (in the left column). The values were not normalized as in Figure 7 or Figure 8, such that the values can be compared between the two columns. For better zooming-in, the scales between the two columns were different. For the cross-validated case, the proportion of variance explained was negative in the baseline time window before 0 ms, which was likely due to the discrepancy between the training and testing subsets. In addition, the maximum proportion of variance explained in the cross-validated case was much smaller ($\approx 2\%$ in the EVC) than the values in the non-data-split case ($\approx 6\%$ in the EVC). However, in the EVC as well as in other ROIs, although the values were smaller, the temporal differential patterns for the three feature sets were similar to those in the non-data-split case. The difference between residual Layer 1 and common components was still present near 300–380 ms in the EVC.

4.5 | The nuisance covariates related to image width, height and aspect ratios

The stimulus images we used had different aspect ratios. In the experiment, we set the longest side of these images to the same length and padded the images into square shapes with gray boundaries. These images were not cropped to the same size for an empirical reason below. The images were chosen initially according to the computer vision model called the never ending image learner (NEIL) (Chen et al., 2013); there were certain features derived from this model that we planned to correlate with neural activity. The NEIL model could be sensitive to all contents in the images, so we did not crop the original images to re-obtain features, but instead, padded all images to the same size.

Later on, at the time of preliminary analysis, we found that the features by NEIL did not explain the neural activity better than CNN models such as AlexNet and did not provide as good interpretability as the layered structure in AlexNet; therefore we used AlexNet instead. We had collected data from half of the participants at the time, so we did not change the images in the middle of the experiment. Instead, in the preprocessing of both the MEG data and the neural network features, we introduced four regressors to capture the width, height, area (which is height $\times$ width) and aspect ratio of the images, and regressed out these nuisance factors.

It is worth noting that these nuisance regressors did explain quite a bit of the variance in the MEG sensor data, especially in the posterior sensors before this preprocessing regression step. See the Figure S5 for the proportion of variance explained for each sensor location (each curve in the lower plot) and the corresponding topology maps at selected time points (in the upper part). Note that the values for the three sensors in each one-magnetometer-two-gradiometer triplet were averaged and plotted as one curve. The colors of the

curves correspond to the sensor location map in the upper-left corner of the plot. For some posterior sensors the proportion of variance reached 14% at 80 ms. This may be because the gray boundary occupied a relatively large visual angle; therefore, they could have evoked large responses in the visual cortex.

4.6 | Noise ceiling of the proportion of variance explained

To evaluate the signal strength in our MEG data, we computed an upper bound of the maximum proportion of variance that could be explained, that is, the “noise ceiling” using a leave-one-participant-out approach. We assume that the best predicting features of the images are not specific to individual participants; they should be an asymptotic ensemble of neural responses to the images from a large group of human participants. This leave-one-participant-out analysis was based on the design that each of the 362 images was presented to all participants; thus for each sensor of each participant at any time point since the stimulus onset, the MEG sensor data, corresponding to the same stimuli near the same time point from all other participants, can be used to predict the data in this sensor. In this case, instead of using predefined features to explain neural responses, we used neural activity from other participants. The features from artificial neural networks should not be able to explain more than this noise ceiling if the participant pool is large enough; in other words, the noise ceiling computed here is an empirical approximation of the upper bound of the variance that can be explained by features of the images.

The implementation of this computation is as follows. Let Participant 1 be the held-out participant, at time point t and for a sensor, we have the MEG data in response to the 362 images, denoted by a 362×1 vector $\mathbf{Y}_{\text{held}}$; correspondingly, we also have the sensor data in response to the 362 images from the remaining 17 participants from all 306 sensors between time points $[t-h, t+h+1]$, which are of dimensions $362 \times 306 \times [2h+1] \times 17$, and can be reshaped to $362 \times (306 \times 17 \times (2h+1))$ and be denoted as $\mathbf{Y}_{\text{remaining}}$. Next, we built linear regression models from $\mathbf{Y}_{\text{remaining}}$ to $\mathbf{Y}_{\text{held}}$, and computed a 10-fold cross-validated proportion of variance in $\mathbf{Y}_{\text{held}}$ explained by $\mathbf{Y}_{\text{remaining}}$. Note that for $\mathbf{Y}_{\text{remaining}}$, we included all sensors to accommodate for the spatial variability of neural activity across participants, and used the window $[t-h, t+h+1]$ to accommodate for temporal variability across participants. For each participant, we ran this leave-one-participant out process, and then took the average of the 10-fold-cross-validated variance explained as the “noise ceiling” for each sensor and each time point. The half time window width $h = 2$ was used at the 100 Hz sampling rate; thus the window was 50 ms long. Note that $\mathbf{Y}_{\text{remaining}}$ has a very large dimension, so in the linear regression, we used PCA to reduce the original dimension to p_0 dimensions, where $p_0 = 6, 10, 20, 40, 80$; we computed the noise ceiling for each sensor at each time point for each p_0 .

Figure S8 shows the maximum noise ceiling over all sensors, for different p_0 s. The maximum noise ceiling over all time points was 35% at 80 ms, when $p_0 = 20$. However, the difference between $p_0 = 6, 10$, and $p_0 = 20$ was small. In later time windows, the noise ceiling

across all sensors decreased quickly to 10% near 150 ms. Figure S9 shows noise ceilings of the individual sensor locations and topology maps, when $p_0 = 20$. Each curve in the lower panel is the average variance explained across three sensors in the one-magnetometer-two-gradiometer triplet in one location. The color of the curves corresponds to the color code of sensors at the upper left. On the upper panel, the topological maps of several time points are shown. The posterior sensors show a higher noise ceiling than other sensors.

Compared with the maximum noise ceiling (35%), the proportion of the variance explained by the 6 dimensional features (*residual Layer 1*, *residual Layer 7* and *common components*) was relatively small (the maximum was 6% to 8% in both the sensor space and ROI-based source-space analysis, without doing data split). This may indicate that there can be features that explain neural data better. As pointed out in the Results section, the local contrast features explained a much larger proportion of variance (with a maximum of about 11%), but these features were likely confounding factors in our comparison of different feature sets, so they were removed.

4.7 | Confounding factors in data-driven experiments

As a coda to our discussion of local contrast as a confounding factor we note that there may exist other image properties that show statistical regularity across naturalistic images. As such, these properties may exhibit significant correlations with neural responses, but be distributed unevenly across the three groups of features (the *common components*, *residual Layer 1*, *residual Layer 7*) we identified in our study. This is a limitation of using naturalistic visual stimuli, where the distribution of features is not an easily manipulable variable. At the same time, in that the visual world typically gives rise to a large dimensional and complicated feature space, it is difficult to form good hypotheses that capture this entire space from scratch. In this context, we view the data-driven exploration of visual processing presented here as an initial step for forming new predictions for future hypothesis-driven experiments.

4.8 | Choice of using AlexNet

We utilized a high-performing CNN model (**AlexNet**; Krizhevsky et al., 2012) to estimate low- and high-level features of the stimulus images. This CNN model was pretrained over 1,000,000's of images for the task of visual object categorization and was among the first “modern” artificial vision models that demonstrated a dramatic improvement over previous models in recognition performance (15.3% top-five error rate). According to our results and other publications from Yamins et al. (2014); Schrimpf et al. (2018), **AlexNet** is also able explain a reasonable amount of the variance in measured neural activity within the visual cortex in both primates and human.

Nevertheless, although our stimuli—naturalistic images of scenes—contained objects that fell into these 1,000 categories, it is unlikely that our participants processed them by only classifying objects. Instead, participants were likely to automatically invoke scene recognition and scene understanding mechanisms during the experiment, although we did not explicitly instruct them to do so. In preliminary

data analyses (not presented here), we used features derived from a network of the same architecture as **AlexNet**, but trained to classify 250 scene categories (Zhou, Lapedriza, Xiao, Torralba, & Oliva, 2014), many of which overlapped with the categories in our stimuli. In this instance, the neural correlation effects we observed were not significantly higher than those with the features from **AlexNet**, which was trained on object, rather than scene, categorization. This finding is consistent with a growing body of results suggesting that the features arising in CNNs trained to perform object categorization (as in **AlexNet**) have good transferability to other visual tasks, as suggested by Yosinski, Clune, Bengio, and Lipson (2014) and Huh, Agrawal, and Efros (2016). One possible explanation is that many of the learned features well characterize broader—task independent—statistical regularities of the visual world. At the same time, scenes contain many object components, and thus scene understanding may benefit from robust mid-level and high-level representations of objects. Supporting this idea, Zhou, Khosla, Lapedriza, Oliva, and Torralba (2014) demonstrated that object detectors emerged in a CNN that was trained on scene classification. Conversely, object categorization requires “partiallying out” variability arising from the different scenes in which a given object may appear—to the extent that particular scenes are consistent across different objects, these scenes may be learned as a route to more effective object invariance. Therefore, it is not entirely surprising that object-category relevant features as instantiated in a CNN can account for neural data during scene processing.

The rapid pace of progress in artificial intelligence also places limitations on even recently run studies using what become quite quickly, less-than-state-of-the-art models. Since the publication of **AlexNet**, deeper, more sophisticated, and higher performing CNNs have been developed for a more diverse set of vision tasks. Although it will be interesting to use features from these new feedforward networks to explain neural data, we hold that it would be more profitable to develop and use non-feedforward neural networks in future work, given the evidence we found of non-feedforward processing in the visual cortex. In particular, such non-feedforward networks should rely on feedback isomorphic to the connectivity structure measured in the primate brain, as well as employing smaller-scale recurrent structures such as lateral inhibition and considering the time scale of information flow within the network (e.g., Nayebi et al., 2018; Yu, Maxfield, & Zelinsky, 2016). In this way, we can compare the dynamics of a high-performing network with spatiotemporal neural activity in the brain so as to better understand the information flow.

4.9 | Limited number of observations and using dimension reduction

As noted, neuroimaging methodologies impose inherent limits on the number of stimuli that may be used: here we presented only 362 scene images. Compared with the high-dimensional features instantiated in **AlexNet**, we did not have a sufficient number of observations to fit large regression models that included all dimensions in the features. In future work, it is important to vastly increase the number of observations (as in Chang, Aminoff, Pyles, Tarr, &

Gupta, 2018), by increasing data collection time or, perhaps, by reducing the number of image repetitions but increasing the signal to noise ratio in the measurement.

We used lower dimensionality for the derived feature sets primarily due to the limitation of data size. One alternative is to rely on representational similarity analysis (RSA) instead of linear regression to test nonlinear dependence, without doing dimensionality reduction. However, for RSA and similar types of nonparametric independence tests (e.g., Gretton, Bousquet, Smola, and Schölkopf (2005)), the testing statistics are mainly useful with permutations to test if significant dependence is present. The statistics themselves are quite sensitive to how many dimensions are used and how redundant the dimensions are in the neural responses and the features. For example, in the feature set, if one dimension that is highly correlated with the neural responses is duplicated, we may obtain higher testing statistics in the RSA. In contrast, a regression analysis with dimension reduction will be more robust to such duplication if the different dimensions in the regressors are orthogonalized. In this sense, using the currently existing RSA-like analysis may not be the best way to compare neural dependence with different sets of features.

4.10 | Limitations of the spatial resolution of MEG

Finally, it is worth pointing out that although our study focuses on joint spatiotemporal patterns of neural activity, the spatial resolution of MEG is inherently limited by the underdetermined nature of the source localization problem. That is, in MEG there are only a limited number of sensors, but there are many more source points in the brain. Moreover, the spatial correlations in the columns of the forward matrix can produce a spatial blurring effect in the reconstructed source solutions. Both these factors add uncertainty in the localization. Although we applied a widely used dSPM method—as well as a sparsity inducing source-space regression method (Yang et al., 2014) as discussed in the Appendix—there is likely a fundamental limit that affects both localization methods. In particular, because of the “underdeterminedness,” we had to exploit some priors, but the true neural activity may violate these assumptions. Consequently, although we have tried our best to make reasonable assumptions while keeping the models tractable, there is some possibility that our source-space results may deviate from the underlying truth. We hope the future development of theories of source localization and the sensor density of MEG (or EEG, see Robinson et al., 2017), as well as experimental work with intracranial recordings in human patients and animals, can further validate our findings.

5 | CONCLUSIONS

In this article, we present detailed spatiotemporal correlation profiles of neural activity using different feature sets derived from the hierarchically-organized layers of an 8-layer CNN—a high-performing vision model for object categorization. Uniquely, we derived three sets of features from the Layer 1 and Layer 7 features of the CNN, which represented the object-category-relevant low-level features (the

common components), the low-level features roughly orthogonal to high-level features (the *residual Layer 1*), and the unique high-level features that were roughly orthogonal to low-level features (the *residual Layer 7*). Our correlation profiles across these feature sets indicated an early-to-late shift from lower-level features to higher-level features and from low-level regions to higher-level regions along the cortical hierarchy, consistent with standard models of feedforward information flow. Moreover, by contrasting the correlation effects of the *common components* and the *residual Layer 1*, we found that the EVC showed a higher and longer correlation effect with the *common components* (the object-category-relevant low-level features) than with the *residual Layer 1* (the low-level features roughly orthogonal to high-level features), in a later time window, possibly in responses to the disappearance of the stimuli. This temporally late, but spatially early, distinction between the two types of low-level features suggests that some non-feedforward processes, such as top-down influences, appears to facilitate the EVC in differentially representing object-category-relevant and less-object-category-relevant information.

ACKNOWLEDGMENTS

This work was supported by 1439237 NSF and RO1 MH64537/MH/NIMH NIH HHS/United States. The author Ying Yang was supported by the Henry L. Hillman Presidential Fellowship at Carnegie Mellon University. We thank Kevin Tan and Austin Marcus for their help in data collection. We also thank the reviewers for constructive suggestions to improve the article.

CONFLICT OF INTERESTS

The first author Ying Yang is an employee at Facebook, Inc. The work in this manuscript was completed entirely before she started working at Facebook.

DATA AVAILABILITY STATEMENT

The preprocessed data that support the findings of this study will be openly available at Figshare (https://figshare.com/articles/MEG_scene_data/7991615). The python scripts that generated the results, as well as links and descriptions to the dataset, are available at https://github.com/YingYang/MEG_Scene.

ORCID

Ying Yang <https://orcid.org/0000-0002-5544-654X>

REFERENCES

- Aminoff, E. M., Toneva, M., Shrivastava, A., Chen, X., Misra, I., Gupta, A., & Tarr, M. J. (2015). Applying artificial vision models to human scene understanding. *Frontiers in Computational Neuroscience*, 9, 8.
- Bar, M., & Aminoff, E. (2003). Cortical analysis of visual context. *Neuron*, 38(2), 347–358.
- Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., et al. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 103(2), 449–454.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57, 289–300.
- Chang, N. C., Aminoff, E. M., Pyles, J. A., Tarr, M. J., & Gupta, A. (2018). *Scaling up neural datasets: A public fMRI dataset of 5000 scenes* (Vol. 18, p. 732). St. Pete Beach, FL: Vision sciences society.
- Chen, X., Shrivastava, A., and Gupta, A. (2013). Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision* 1409–1416.
- Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, 153, 346–358.
- Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016b). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6, 1–13.
- Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016c). Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition. *arXiv preprint arXiv:1601.02970*.
- Cichy, R. M., Pantazis, D., & Oliva, A. (2016). Similarity-based fusion of MEG and fMRI reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cerebral Cortex*, 26(8), 3563–3579.
- Clarke, A., Devereux, B. J., Randall, B., & Tyler, L. K. (2014). Predicting the time course of individual objects with MEG. *Cerebral Cortex*, 25(10), 3602–3612.
- Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, 29(3), 162–173.
- Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., & Halgren, E. (2000). Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, 26(1), 55–67.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. IEEE.
- DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8), 333–341.
- Engel, S. A., Glover, G. H., & Wandell, B. A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cerebral Cortex*, 7(2), 181–192.
- Epstein, R., Harris, A., Stanley, D., & Kanwisher, N. (1999). The parahippocampal place area: Recognition, navigation, or encoding? *Neuron*, 23(1), 115–125.
- Fellbaum, C. (1998). *Wordnet: An electronic lexical database*. Bradford book.
- Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1), 1–47.
- Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: Automated labeling of neuro-anatomical structures in the human brain. *Neuron*, 33(3), 341–355.
- 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(7), 974–981.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).
- Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, 86, 446–460.

- Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory* (pp. 63–77). Berlin, Heidelberg: Springer.
- Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzhak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron*, 24(1), 187–203.
- Hamalainen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., & Lounasmaa, O. V. (1993). Magnetoencephalography—theory, instrumentation, to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65, 414–487.
- Hedrich, T., Pellegrino, G., Kobayashi, E., Lina, J.-M., & Grova, C. (2017). Comparison of the spatial resolution of source imaging techniques in high-density EEG and MEG. *NeuroImage*, 157, 531–544.
- Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1), 215–243.
- Huh, M., Agrawal, P., and Efros, A. A. (2016). What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*.
- Ito, M., & Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area v2 of macaque monkeys. *The Journal of Neuroscience*, 24(13), 3313–3324.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM.
- Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 10(11), e1003915.
- Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9), 1464–1480.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Leeds, D. D., Seibert, D. A., Pyles, J. A., & Tarr, M. J. (2013). Comparing visual representations across human fMRI and computational vision. *Journal of Vision*, 13(13), 25.
- Liang, Z., Shen, W., Sun, C., & Shou, T. (2008). Comparative study on the offset responses of simple cells and complex cells in the primary visual cortex of the cat. *Neuroscience*, 156(2), 365–373.
- Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1), 177–190.
- Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., DiCarlo, J. J., and Yamins, D. L. (2018). Task-driven convolutional recurrent models of the visual system. In *Advances in Neural Information Processing Systems*, pages 5295–5306.
- Nestor, A., Vettel, J. M., & Tarr, M. J. (2008). Task-specific codes for face recognition: How they shape the neural representation of features for detection and individuation. *PLoS One*, 3(12), e3978.
- Robinson, A. K., Venkatesh, P., Boring, M. J., Tarr, M. J., Grover, P., & Behrmann, M. (2017). Very high density EEG elucidates spatiotemporal aspects of early visual processing. *Scientific Reports*, 7(1), 16248.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252.
- Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.
- Ségonne, F., Dale, A., Busa, E., Glessner, M., Salat, D., Hahn, H., & Fischl, B. (2004). A hybrid approach to the skull stripping problem in MRI. *NeuroImage*, 22(3), 1060–1075.
- Sejnowski, T. J., Churchland, P. S., & Movshon, J. A. (2014). Putting big data to good use in neuroscience. *Nature Neuroscience*, 17(11), 1440–1441.
- Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19(1), 109–139.
- Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582), 520–522.
- Wasserman, L. (2010). *All of statistics: A concise course in statistical inference*. Springer-Verlag, New York: Springer Publishing Company Incorporated.
- Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE.
- Xu, Y., Sudre, G. P., Wang, W., Weber, D. J., & Kass, R. E. (2011). Characterizing global statistical significance of spatiotemporal hot spots in magnetoencephalography/electroencephalography source space via excursion algorithms. *Statistics in Medicine*, 30(23), 2854–2866.
- Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23), 8619–8624.
- Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3), 356–365.
- Yang, Y. (2017). *Source-space analyses in MEG/EEG and applications to explore spatio-temporal neural dynamics in human vision*. (PhD dissertation), Carnegie Mellon University.
- Yang, Y., Tarr, M. J., & Kass, R. E. (2014). Estimating learning effects: A short-time fourier transform regression model for MEG source localization. In *Lecture notes on artificial intelligence: MLINI 2014: Machine learning and interpretation in neuroimaging*. Montreal, Canada: Springer.
- Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in neural information processing systems*, 3320–3328.
- Yu, C.-P., Maxfield, J., & Zelinsky, G. J. (2016). Generating the Features for Category Representation using a Deep Convolutional Neural Network. *Vision Sciences Society*, 16, 1161876.
- Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* pages 818–833. Springer.
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2014). Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*.
- Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Yang Y, Tarr MJ, Kass RE, Aminoff EM. Exploring spatiotemporal neural dynamics of the human visual cortex. *Hum Brain Mapp*. 2019;40:4213–4238. <https://doi.org/10.1002/hbm.24697>