Deep Residual Network Reveals a Nested Hierarchy of Distributed Cortical Representation for Visual Categorization

Haiguang Wen[2,3], Junxing Shi[2,3], Wei Chen[4], Zhongming Liu[*1,2,3]

[1]Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN, USA

[2]School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA

[3]Purdue Institute for Integrative Neuroscience, Purdue University, West Lafayette, IN, USA

[4]Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota Medical School, Minneapolis, MN, USA

***Correspondence**

Zhongming Liu, PhD

Assistant Professor of Biomedical Engineering

Assistant Professor of Electrical and Computer Engineering

College of Engineering, Purdue University

206 S. Martin Jischke Dr.

West Lafayette, IN 47907, USA

Phone: +1 765 496 1872

Fax: +1 765 496 1459

Email: zmliu@purdue.edu

Abstract

What enables humans to readily recognize visual objects is attributed to how the brain extracts, represents, and organizes object information from visual input. To understand this process, we used a deep residual network as a hierarchical computational model of visual categorization to predict cortical representation of natural vision. We trained and tested such a predictive encoding model with functional magnetic resonance imaging data from human subjects watching hours of natural video clips, and verified its ability to predict cortical responses to novel images, objects, and categories. We further used the so trained encoding model to synthesize cortical responses to 64,000 visual objects from 80 categories, revealing hierarchical, distributed, and overlapping cortical representations of categories. Such category representations covered both the ventral and dorsal pathways, reflected multiple levels and domains of visual features, and preserved semantic relationships between categories. In the scale of the entire visual cortex, category representations were modularly organized into three categories: biological objects, non-biological objects, and background scenes. In a smaller scale specific to each module, category representation further revealed sub-modules for finer categorization, e.g. biological objects were categorized into terrestrial animals, aquatic animals, humans, and plants. Such nested spatial and representational hierarchies were attributable to different levels of category information to varying degrees. These findings suggest that increasingly more specific category information is represented by cortical patterns in progressively finer spatial scales – an important principle for the brain to categorize visual objects in various levels of abstraction.

Keywords: natural vision, object categorization, deep learning, encoding model

Significance Statement

This study uses a deep neural network to model how the brain extracts and represents the information from visual objects for efficient and flexible categorization. Results show that the model can predict widespread cortical responses when humans are viewing natural video clips, and generate patterns of cortical responses to tens of thousands of objects from 80 categories. Analysis of these response patterns reveals that the brain organizes distributed, overlapping, and property-based category representations into a nested hierarchy. Increasingly more specific category information is represented by cortical patterns in progressively finer spatial scales. This nested hierarchy may be a fundamental principle for the brain to categorize visual objects in various levels of specificity.

Introduction

The visual cortex performs rapid categorization of complex and diverse visual patterns or objects. This ability is attributable to hierarchical neural computation and representation of category information (Van Essen et al., 1992; Yamins and DiCarlo, 2016). In particular, the ventral temporal cortex contains topologically organized maps of object representations (Grill-Spector and Weiner, 2014), spanning a high-dimensional space (Haxby et al., 2011) while being invariant against changes in low-level visual properties (Quiroga et al., 2005; DiCarlo and Cox, 2007). Evidence shows that category representations also exist in the dorsal stream (Chao and Martin, 2000; Bracci and de Beeck, 2016; Freud et al., 2016) or even beyond (Gallese and Lakoff, 2005), likely reflecting non-visual attributes needed for category dependent actions (Martin, 2007). It has thus been proposed that distributed cortical networks extract and represent category information for robust, efficient, and flexible visual categorization in multiple levels of abstraction (Chao et al., 1999; Haxby et al., 2001).

However, the computational understanding of distributed neural coding is still limited. Questions are unresolved as to how information is encoded in distributed patterns (Haxby et al., 2001), how object knowledge emerges from lower-level visual features (Yamins and DiCarlo, 2016), and how cortical representations share and differ across categories (Grill-Spector and Weiner, 2014). Answering such questions requires a fully accessible computational model of hierarchical cortical processing for visual categorization (Riesenhuber and Poggio, 1999), and mapping representations of as many categories as possible in a huge, if not infinite, dimension of object domains (Huth et al., 2012). These challenges and requirements may be met by recent advances in deep neural networks (DNN) (LeCun et al., 2015) – a type of artificial neural

networks built with conceptually similar architecture and computing principle as the brain itself (Yamins and DiCarlo, 2016).

Recent studies show that convolutional neural networks offer hierarchical representations of any visual input to be able to model and predict cortical responses to natural picture (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Güçlü and van Gerven, 2015b; Cichy et al., 2016; Eickenberg et al., 2017) or video (Güçlü and van Gerven, 2015a; Wen et al., 2016) stimuli. The predictive power of such network models is high and robust in the entire visual cortex (Wen et al., 2016), rendering them more favorable than other models that only account for either the lowest (Kay et al., 2008; Nishimoto et al., 2011) or highest (Huth et al., 2012) level in the visual hierarchy. The DNN-based predictive model can be applied to novel (unseen) visual stimuli (Yamins et al., 2014; Güçlü and van Gerven, 2015b; Wen et al., 2016; Eickenberg et al., 2017). It thus enables to simulate the cortical representations of a large number of visual objects and categories (Wen et al., 2016; Eickenberg et al., 2017), far beyond what is attainable experimentally (Kiani et al., 2007; Mahon et al., 2009; Kourtzi and Connor, 2011; Mur et al., 2012; Naselaris et al., 2012).

Extending from recent studies (Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and van Gerven, 2015b; Cichy et al., 2016; Wen et al., 2016; Eickenberg et al., 2017), we used a deep residual network (ResNet) (He et al., 2016) to define, train, and test a generalizable, predictive, and hierarchical model of natural vision by using extensive functional magnetic resonance imaging (fMRI) data from humans watching >10 hours of YouTube videos. Taking this predictive model as a "virtual" fMRI scanner, we synthesized the cortical response patterns with 64,000 natural pictures including objects from 80 categories, and mapped category representations with high-throughput. We analyzed and compared the cortical representational

similarity among categories against their semantic relationships, and separated the contributions from different levels of visual information to object categorization. A primary focus here was on testing a hypothesis that the brain uses nested spatial and representational hierarchies to support multi-level visual categorization. Object representations in large to small scales support coarse to fine categorization (Grill-Spector and Weiner, 2014).

Materials and Methods

Experimental data

We used and extended the experimental data from our previous study (Wen et al., 2016). Briefly, the data included the fMRI scans from three healthy subjects (Subject 1, 2, 3) when watching natural videos. For each subject, the video-fMRI data were split into two independent datasets: one for training the encoding model and the other for testing it. For Subject 2 & 3, the training movie included 2.4 hours of videos; the testing movie included 40 minutes of videos; the training movie was repeated twice, and the testing movie was repeated ten times. For Subject 1, the training movie included not only those videos presented to Subject 2 and 3, but also 10.4 hours of new videos. The movie stimuli included a total of ~9,300 video clips manually selected from *YouTube* (https://www.youtube.com), covering a variety of real-life visual experiences. All video clips were concatenated in a random sequence and separated into 8-min sessions. Every subject watched each session of videos (field of view: $20.3^{o}\times20.3^{o}$) through a binocular goggle with the eyes fixating at a central cross ($0.8^{o}\times0.8^{o}$). During each session, whole-brain fMRI scans were acquired with 3.5 mm isotropic resolution and 2 s repetition time in a 3-T MRI system. The volumetric fMRI data were preprocessed and co-registered onto a standard cortical

surface template (Glasser et al., 2013). More details about the movie stimuli, data preprocessing and acquisition are described elsewhere (Wen et al., 2016).

Deep residual network

In line with previous studies (Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and van Gerven, 2015b; Cichy et al., 2016; Wen et al., 2016; Eickenberg et al., 2017; Horikawa and Kamitani, 2017), a feedforward deep neural network (DNN) was used to model the cortical representations of natural visual stimuli. Here, we used a specific version of the DNN known as the deep residual network (ResNet), which had been pre-trained to categorize natural pictures with the state-of-the-art performance (He et al., 2016). In the ResNet, 50 hidden layers of neuron-like computational units were stacked into a bottom-up hierarchy. The first layer encoded location and orientation-selective visual features, whereas the last layer encoded semantic features that supported categorization. The layers in between encoded increasingly complex features through 16 residual blocks; each block included three successive layers and a shortcut directly connecting the input of the block to the output of the block (He et al., 2016). Compared to the DNNs in prior studies (Cadieu et al., 2014; Yamins et al., 2014; Güçlü and van Gerven, 2015b; Cichy et al., 2016; Wen et al., 2016; Horikawa and Kamitani, 2017), the ResNet was much deeper and defined more fine-grained hierarchical visual features. The ResNet could be used to extract feature representations from any input image or video frame by frame. Passing an image into the ResNet yielded an activation value at each unit. Passing a video yielded an activation time series at each unit as the fluctuating representation of a given visual feature in the video.

Encoding models

For each subject, we trained an encoding model to predict each voxel's fMRI response to any natural visual stimuli (Naselaris et al., 2011), using a similar strategy as previously explored (Güçlü and van Gerven, 2015b; Wen et al., 2016; Eickenberg et al., 2017). The voxel-wise encoding model included two parts: the first part was nonlinear, converting the visual input from pixel arrays into representations of hierarchical features through the ResNet; the second part was linear, projecting them onto each voxel's fMRI response. The encoding model used the features from 18 hidden layers in the ResNet, including the first layer, the last layer, and the output layer for each of the 16 residual blocks. For video stimuli, the time series extracted by each unit was standardized (i.e. remove the mean and normalize the variance), and convolved with a canonical hemodynamic response function (HRF) with the peak response at 4s, and then down-sampled to match the sampling rate of fMRI.

The feature dimension was reduced by applying principle component analysis (PCA) first to each layer and then to all layers in ResNet. The principal components of each layer were a set of orthogonal vectors that explained >99% variance of the layer's feature representations given the training movie. The layer-wise dimension reduction was expressed as Eq. (1).

$$\boldsymbol{f}_l(\mathbf{x}) = \boldsymbol{f}_l^o(\mathbf{x})\mathbf{B}_l \qquad (1)$$

where $\boldsymbol{f}_l^o(\mathbf{x})$ $(1\times p_l)$ is the original feature representation from layer l given a visual input $\mathbf{x}$, $\mathbf{B}_l$ $(p_l\times q_l)$ consists of unitary columnar vectors that represented the principal components for layer l, $\boldsymbol{f}_l(\mathbf{x})$ $(1\times q_l)$ is the feature representation after reducing the dimension from p_l to q_l.

Following the layer-wise dimension reduction, the feature representations from all layers were further reduced by using PCA to retain >99% variance across layers. The final dimension reduction was implemented as Eq. (2).

$$\boldsymbol{f}(\mathbf{x}) = \boldsymbol{f}_{1:L}(\mathbf{x})\mathbf{B}_{1:L} \qquad (2)$$

where $\boldsymbol{f}_{1:L}(\mathbf{x}) = \left[\frac{f_1(\mathbf{x})}{\sqrt{p_1}}, \ldots, \frac{f_L(\mathbf{x})}{\sqrt{p_L}}\right]$ is the feature representation concatenated across L layers, $\mathbf{B}_{1:L}$ consists of unitary principal components of the layer-concatenated feature representations of the training movie, and $\boldsymbol{f}(\mathbf{x})$ $(1\times k)$ is the final dimension-reduced feature representation.

For the second part of the encoding model, a linear regression model was used to predict the fMRI response $r_v(\mathbf{x})$ at voxel v evoked by the stimulus $\mathbf{x}$ based on the dimension-reduced feature representation $\boldsymbol{f}(\mathbf{x})$ of the stimulus, as expressed by Eq. (3).

$$r_v(\mathbf{x}) = \boldsymbol{f}(\mathbf{x})\,\mathbf{w}_v + \varepsilon_v \qquad (3)$$

where $\mathbf{w}_v$ is a columnar vector of regression coefficients specific to voxel v, and ε_v is the error term. As shown in Eq. (4), L_2-regularized least-squares estimation was used to estimate $\mathbf{w}_v$ given the data during the training movie (individual frames were indexed by $i = 1, \cdots, N$), where the regularization parameter was determined based on nine-fold cross-validation.

$$\widehat{\mathbf{w}}_v = \arg\min_{\mathbf{w}_v} \frac{1}{N}\sum_{i=1}^{N}(r_v(\mathbf{x}_i) - \boldsymbol{f}(\mathbf{x}_i)\,\mathbf{w}_v)^2 + \lambda\|\,\mathbf{w}_v\|_2^2 \qquad (4)$$

After the above training, the voxel-wise encoding models were evaluated for their ability to predict the cortical responses to the novel testing movie (not used for training). The prediction accuracy was quantified as the temporal correlation (r) between the predicted and observed fMRI responses at each voxel given the testing movie. Since the testing movie included five distinct sessions, the prediction accuracy was evaluated separately for each session, converted to the z-

score using the Fisher's z-transform, and then averaged across sessions. The significance of the voxel-wise prediction accuracy was evaluated with a block-permutation test (Adolf et al., 2014) (corrected at false discovery rate (FDR) $q < 0.01$), as used in our prior study (Wen et al., 2016).

We tested whether the deeper ResNet outperformed the shallower AlexNet (Krizhevsky et al., 2012) in predicting cortical responses to natural movies, taking the latter as the benchmark given its state-of-the-art encoding performance in prior studies (Güçlü and van Gerven, 2015b; Cichy et al., 2016; Wen et al., 2016). For this purpose, we trained and tested similar encoding models based on the AlexNet with the same analysis of the same dataset. We compared the prediction accuracy (as the z value) between ResNet and AlexNet for regions of interest (ROIs) defined in an existing cortical parcellation (Glasser et al., 2016), and further evaluated the statistical significance of their difference using a paired t-test ($p<0.001$) across all voxels within each ROI.

Human-face representations with encoding models and functional localizer

The ResNet-based encoding models were further used to simulate cortical representations of human faces, in comparison with the results obtained with a functional localizer applied to the same subjects. To simulate the cortical "face" representation, 2,000 human-face pictures were obtained by Google Image search. Each of these pictures was input to the voxel-wise encoding model, simulating a cortical response map as if it were generated when the subject was actually viewing the picture, as initially explored in previous studies (Wen et al., 2016; Eickenberg et al., 2017). The simulated response maps were averaged across all the face pictures, synthesizing the cortical representation of human face as an object category.

To validate the model-synthesized "face" representation, a functional localizer (Fox et al., 2009) was used to experimentally map the cortical face areas on the same subjects. Each subject participated in three sessions of fMRI with a randomized block-design paradigm. The paradigm included alternating ON-OFF blocks with 12s per block. During each ON block, 15 pictures (12 novel and 3 repeated) from one of the three categories (face, object, and place) were shown for 0.5s per each picture with a 0.3s interval. The ON blocks were randomized and counter-balanced across the three categories. Following the same preprocessing as for the video-fMRI data, the block-design fMRI data were analyzed with a general linear model (GLM) with three predictors, i.e. face, object, and place. Cortical "face" areas were localized by testing the significance of a contrast (face>object and face > place) with $p<0.05$ and Bonferroni correction.

Synthesizing cortical representations of different categories

Beyond the proof of concept with human faces, the similar strategy was also extended to simulate the cortical representations of 80 categories through the ResNet-based encoding models. The category labels were shown in Table 1. These categories were mostly covered by the video clips used for training the encoding models. For each category, 800 pictures were obtained by Google Image search with the corresponding label, and were visually inspected to replace any exemplar that belonged to more than one category. The cortical representation of each category was generated by averaging the model-simulated response map given every exemplar within the category.

Category selectivity

Following the above analysis, cortical representations were compared across categories to quantify the category selectivity of various locations and ROIs. For each voxel, its selectivity to category i against other categories i^c was quantified with Eq. (5), as previously suggested (Afraz et al., 2006).

$$d_i' = \frac{\bar{r}_i - \bar{r}_{i^c}}{\sqrt{(\sigma_i^2 + \sigma_{i^c}^2)/2}} \qquad (5)$$

where $\bar{r}_i$ and σ_i^2 are the mean and variance of the responses to the exemplars in category i, and $\bar{r}_{i^c}$ and $\sigma_{i^c}^2$ were counterparts to all exemplars in other categories i^c. Irrespective of any specific category, the general category-selectivity for each voxel was its maximal d' index among all categories, i.e. $d' = \max_i\{d_i'\}$. A d' index of zero suggests non-selectivity to any category, and a higher d' index suggests higher category-selectivity. The category selectivity of any given voxel was also inspected by listing the categories in a descending order of their representations at the voxel. We also obtained the ROI-level category selectivity by averaging the voxel-wise selectivity across voxels and subjects. ROIs were defined in an existing cortical parcellation (Glasser et al., 2016).

Categorical similarity and modularity in cortical representation

To reveal how the brain organizes categorical information, we assessed the similarity (i.e. spatial correlation) in cortical representations between categories. Based on such inter-category similarity, individual categories were grouped into clusters using k-means clustering (MacQueen, 1967). The goodness of clustering was measured as the modularity index, which quantified the inter-category similarities within the clusters relative to those regardless of the clusters (Gómez et al., 2009).

The similarity in cortical representation between different categories was compared with their similarity in semantic meaning. The semantic similarity between categories was evaluated as the Leacock-Chodorow similarity (Leacock and Chodorow, 1998) between the corresponding labels based on their relationships defined in the WordNet (Fellbaum, 1998) – a directed graph of words (as the nodes) and their *is-a* relationships (as the edges). The correlation between the cortical and semantic similarities was evaluated across all pairs of categories.

Layer-wise contribution to cortical categorical representation

We also asked which levels of visual information contributed to the modular organization of categorical representations in the brain. To answer this question, the cortical representation of each category was dissected into multiple levels of representations, each of which was attributed to one single layer of features. For a given category, the features extracted from every exemplar of this category were kept only for one layer in the ResNet, while setting to zeros for all other layers. Through the voxel-wise encoding model, the single-layer visual features were projected onto a cortical map that only represented a certain level of visual information shared in the given category. The similarity and modularity in cortical representations of individual categories were then re-evaluated as a function of the layer in the ResNet. The layer with the highest modularity index contributed the most to the modular organization in cortical categorical representation. The features encoded by this layer were visualized for more intuitive understanding of the types of visual information underlying the modular organization. The feature visualization was based on an optimization-based technique (Yosinski et al., 2015). Briefly, to visualize the feature encoded by a single unit in the ResNet, the input to the ResNet was optimized to iteratively maximize the

output from this unit, starting from a Gaussian random pattern. Four optimized visualizations were obtained given different random initialization.

Categorical representation in nested hierarchical spatial scales

Considering object categories were defined hierarchically in semantics (Fellbaum, 1998), we asked whether there were spatial and representational hierarchies underlying the hierarchy of categorization – a hypothesis proposed in (Grill-Spector and Weiner, 2014). More specifically, we tested whether the representational similarity and distinction in a larger spatial scale gave rise to a coarser level of categorization, whereas the representation in a smaller spatial scale gave rise to a finer level of categorization. To do so, we first examined the category representation in the scale of the entire visual cortex predictable by the encoding models, and clustered the categories into multiple modules by using the modularity analysis of the representational similarity in this large scale (see ***Categorical similarity and modularity in cortical representation***). The resulting modules of categories were compared with the superordinate-level semantic categories. Then, we focused on a finer spatial scale specific to the regions where category representations overlapped within each module in contrast to 50,000 random and non-selective objects ($p<0.01$, two-sample t-test, Bonferroni correction). Given the spatial similarity of category representation in this finer scale, we defined sub-modules within each module using the same modularity analysis as for the large-scale representation. The sub-modules of categories were compared and interpreted against semantic categories in a finer level.

Results

ResNet predicted widespread cortical responses to natural visual stimuli

In line with recent studies (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Güçlü and van Gerven, 2015b; Cichy et al., 2016; Wen et al., 2016; Eickenberg et al., 2017), we used a deep convolutional neural network to establish predictive models of the cortical fMRI representations of natural visual stimuli. Specifically, we used the ResNet – a deep residual network pre-trained for computer vision (He et al., 2016), with a much deeper architecture to yield more fine-grained layers of visual features than otherwise similar but shallower networks, e.g. AlexNet (Krizhevsky et al., 2012), explored in prior studies (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Güçlü and van Gerven, 2015b; Cichy et al., 2016; Wen et al., 2016; Eickenberg et al., 2017; Horikawa and Kamitani, 2017). From any visual stimuli, the ResNet-extracted features jointly predicted the fMRI response through a voxel-wise linear regression model (see ***Encoding models*** in **Methods and Materials**). This encoding model was trained with a large amount of fMRI data during a training movie (12.8 hours for Subject 1, and 2.4 hours for Subject 2, 3), and tested with an independent testing movie (40 minutes).

The encoding accuracy (i.e. the correlation between the predicted and measured fMRI signals during the testing movie) was overall high ($r = 0.43 \pm 0.14$, 0.36 ± 0.12, and 0.37 ± 0.11 for Subject 1, 2 and 3, respectively) and statistically significant (permutation test, corrected at FDR $q<0.01$) throughout the visual cortex in every subject (Fig. 1.a). The encoding accuracy was comparable among the higher-order ventral-stream areas, e.g. fusiform face area (FFA) and parahippocampal place area (PPA), as well as early visual areas, e.g. V1, V2, V3, and V4 (Fig. 1.b), whereas it was relatively lower at such dorsal-stream areas as lateral intraparietal area (LIP), frontal eye fields (FEF), parietal eye fields (PEF), but not the middle temporal area (MT)

(Fig. 1.b). The prediction accuracy was consistently higher with (the deeper) ResNet than with (the shallower) AlexNet (Fig. 1.b). These results suggest that the ResNet-based voxel-wise encoding models offer generalizable computational accounts for the complex and nonlinear relationships between natural visual stimuli and cortical responses at widespread areas involved in various levels of visual processing.

Encoding models predicted cortical representations of various object categories

As explored before (Wen et al., 2016; Eickenberg et al., 2017), the voxel-wise encoding models constituted a high-throughput computational workbench to synthesize cortical activations with a large number of natural pictures, not realistically attainable with experimental approaches. Here, we used this strategy to predict the pattern of cortical activation with each of the 64,000 natural pictures from 80 categories with 800 exemplars per category (see ***Synthesizing cortical representations of different categories*** in **Methods and Materials**). By averaging the predicted activation maps across all exemplars of each category, the common cortical activation within this category was obtained to report its cortical representation.

For example, averaging the predicted responses to various human faces revealed the category-wide cortical representation of the "face" invariant of low-level visual features, e.g. the color, position, and perspective (Fig. 2.b). Such a model-simulated "face" representation was consistent with the fMRI-mapping result obtained with a block-design functional localizer that contrasted face vs. non-face pictures (Fig. 2.a). In a similar manner, cortical representations of all 80 categories were mapped (Fig. 3). The resulting category representations were not only along the ventral stream, but also along the dorsal stream albeit with relatively lower amplitudes and a smaller extent.

For each voxel, the model-predicted response as a function of category was regarded as the voxel-wise profile of categorical representation. The category selectivity – a measure of how a voxel was selectively responsive to one category relative to others (Afraz et al., 2006), varied considerably across cortical locations (Fig. 4.a). Voxels with higher category selectivity were clustered into discrete regions including the bilateral PPA, FFA, lateral occipital (LO) area, the temporo-parietal junction (TPJ), as well as the right superior temporal sulcus (STS) (Fig. 4.a). The profile of categorical representation listed in a descending order (Fig. 4.b), showed that FFA, OFA, and pSTS were selective to humans or animals (e.g. man, woman, monkey, cat, lion); PPA was highly selective to places (e.g. kitchen, office, living room, corridor); the ventral visual complex (VVC) was selective to man-made objects (e.g. cellphone, tool, bowl, car). In general, the ventral stream tended to be more category-selective than early visual areas (e.g. V1, V2, V3) and dorsal-stream areas (e.g. MT, LIP) (Fig. 4.c).

Distributed, overlapping, and modular representations of categories

Although some ventral-stream areas (e.g. PPA and FFA) were highly (but not exclusively) selective to a certain category, no category was represented by any single region alone (Fig. 3). As suggested previously (Haxby et al., 2001), object categories were represented by distributed and overlapping networks (see examples in Fig. 5). In the scale of the nearly entire visual cortex as predictable by the encoding models (Fig. 1.a), the spatial correlations in cortical representation between distinct categories were shown as a representational similarity matrix (Fig 6.a). This matrix revealed a modular organization (modularity Q=0.35), by which categories were clustered into three superordinate-level modules (Fig. 6.a, left). The categories being clustered based on their cortical representations exhibited a similarly modular pattern in terms of their semantic

similarity (Fig. 6.a, middle), measured as the LCH similarity between the corresponding labels in WordNet (Leacock and Chodorow, 1998). Interestingly, the similarity in cortical representation between categories was highly correlated with their semantic similarity (Fig. 6.a, right), suggesting that categories with more similar cortical representations tend to bear more closely related semantic meanings.

The representational modules in the entire visual cortex revealed coarse categories that seemed reasonable. The first module included non-biological objects, e.g. airplane, bottle and chair; the second module included biological objects, e.g. humans, animals, and plants; the third module included places and scenes (Fig. 6.b). The cortical representation averaged within each module revealed the general cortical representations of non-biological and biological objects, and background scenes (Fig. 6.b). As shown in Fig. 6.b, non-biological objects were represented by activations in bilateral sub-regions of ventral temporo-occipital cortex (e.g. VVC); biological objects were represented by activations in the lateral occipital cortex and part of the inferior temporal cortex (e.g. FFA) but deactivations in parahippocampal cortex (e.g. PPA); background scenes were represented by activations in PPA but deactivations in the lateral occipital complex, partly anti-correlated with the activations with biological objects.

Mid-level visual features accounted for basic-level categorization

Which levels of visual features accounted for such a modular organization were revealed by examining the representational similarity and modularity as attributed to the features extracted by each layer in the ResNet (see ***Layer-wise contribution to cortical categorical representation***). Fig. 7.a (left) shows the inter-category representational similarity given the layer-wise features, thus decomposing the modular organization in Fig. 6.a by layers. The layer-wise modularity in

cortical representation emerged progressively, being entirely absent given the 1st layer, showing noticeable three modules from the 13th layer, and reaching the maximum at the 31st layer (Fig. 7.a).

To gain intuition about the types of visual information from the 31st layer, the features encoded by individual units in this layer were visualized (see ***Layer-wise contribution to cortical categorical representation***). Fig. 7.b illustrates the visualizations of some example features, showing shapes or patterns (both 2-D and 3-D), animal or facial parts (e.g. head and eye), environmental components (e.g. house and mountain). Beyond these examples, other features were of similar types. Therefore, the mid-level features that depict object shapes or parts are modularly organized by their distributed cortical representations, supporting the superordinate-level categorization.

More specific categories were modularly organized in finer scales

We further asked whether the similar modular organization could be extended to a lower level of categorization. That is, whether object representations were modularly organized within each superordinate-level module. For this purpose, we confined the scope of analysis from the whole visual cortex (Fig. 1.a) to finer spatial scales highlighted by co-activation patterns within biological objects, non-biological objects, or background scenes (Fig. 8.a). For example, within the regions where biological objects were represented (Fig. 8.a, top), the representational patterns were further clustered into four sub-modules: terrestrial animals, aquatic animals, plants, and humans (Fig. 8.b, top). Similarly, the fine-scale representational patterns of background scenes were clustered into two sub-modules corresponding to artificial (e.g. bedroom, bridge, restaurant) and natural scenes (e.g. falls, forest, beach) (Fig. 8, middle). However, non-biological objects

showed a much less degree of modularity in cortical representation; the two modules did not bear any reasonable conceptual distinction (Fig. 8, bottom).

We also evaluated the layer-wise contribution to the fine-scale representational similarity and modularity. For biological objects, the modularity index generally increased from the lower to higher layer, reaching the maximum at the highest layer (Fig. 9.a, top). Note that the highest layer encoded the most abstract and semantically relevant features, whose visualizations revealed the entire objects or scenes (Fig. 9.b) rather than object or scenic parts (Fig. 7.b). In contrast, the modularity index reached the maximum at the 28th layer for background scenes (Fig. 9.a, middle), but was relatively weak and less layer-dependent for non-biological objects (Fig. 9.a, bottom).

Discussion

This study demonstrates a high-throughput computational strategy to characterize hierarchical, distributed, and overlapping cortical representations of visual objects and categories. Results suggest that information about visual-object category entails multiple levels and domains of features represented by distributed cortical patterns in both ventral and dorsal pathways. Categories with similar cortical representations are more related in semantics. In a large scale of the entire visual cortex, object representations are modularly organized into three superordinate categories (biological objects, non-biological objects, and background scenes). In a finer scale specific to each module, category representation reveals sub-modules for finer categorization (e.g. biological objects are categorized into terrestrial animals, aquatic animals, plants, and humans). The cortical organization of category representation is attributed to middle to high levels of category information to a varying degree. These findings support a nested hierarchy in distributed cortical representation for visual categorization: increasingly more

specific category information is represented by distinct cortical patterns in progressively finer spatial scales (Grill-Spector and Weiner, 2014), enabling the brain to identify, relate, and separate objects in various levels of abstraction.

ResNet predicts and dissects cortical representations of categories

Central to this study is the use of the categorization-driven deep ResNet for synthesizing the cortical representations of thousands of natural visual objects from many categories. This strategy has a much higher throughput for sampling a virtually infinite object or category space (Wen et al., 2016; Eickenberg et al., 2017), compared to prior studies that are limited to fewer categories with much fewer exemplars per category (Kiani et al., 2007; Mahon et al., 2009; Kourtzi and Connor, 2011; Mur et al., 2012; Naselaris et al., 2012). The sample size should be further extendable, since the ResNet-based encoding models account for the relationships between cortical responses and hierarchical and invariant visual features. Such features are finite and generalizable to different and new natural images, objects, and categories which the models have not been explicitly trained with. The model predictions are reasonably accurate and consistent with experimentally observed cortical responses (Fig. 1.a) and object representations (Fig. 2). The encoding accuracy may be further improved given an even larger and more diverse video-fMRI dataset to train the model, and a more biologically relevant deep neural net that better matches the brain and better performs in computer-vision tasks (Yamins et al., 2014). In this sense, the encoding models in this study are based on so far largest video-fMRI training data from single subjects; ResNet also outperforms AlexNet in categorizing images (Krizhevsky et al., 2012; He et al., 2016) and predicting the brain (Fig. 1.b). The encoding models reported here are thus arguably more powerful in predicting and mapping hierarchical cortical representations

across the entire visual cortex (Fig. 1), compared to conceptually similar models in prior studies (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Güçlü and van Gerven, 2015b; Cichy et al., 2016; Wen et al., 2016; Eickenberg et al., 2017).

What is also advantageous is that ResNet decomposes category information into multiple layers of features progressively emerging from lower to higher levels. As such, ResNet offers a computational account of hierarchical cortical processing for categorization, yielding quantitative description of every object or category in terms of different layers of visual features. Mapping the layer-wise features from the ResNet onto the brain helps to address what drives the cortical organization of object knowledge and supports various levels of categorization.

Distributed and overlapping cortical category representations

Our results support the notion that visual-object categories are represented by distributed and overlapping cortical patterns (Haxby et al., 2001) rather than clustered regions (Kanwisher et al., 1997; Epstein and Kanwisher, 1998; Peelen and Downing, 2005). Given this notion, the brain represents a category not as a single entity but a set of defining attributes that span multiple domains and levels of object knowledge. Different objects may bear overlapping representational patterns that are both separable and associable, allowing them to be recognized as one category in a particular level, but as different categories in another level. For example, a lion and a shark are both animals but can be more specifically categorized as terrestrial and aquatic animals, respectively. The distributed and overlapping object representations, as weighted spatial patterns of attribute-based representations (Martin, 2007), constitute an essential principle underlying the brain's capacity for multi-level categorization.

Category representations, although distributed in general, may become highly selective at

spatially clustered regions (Kanwisher et al., 1997; Epstein and Kanwisher, 1998; Peelen and Downing, 2005). The category-selective regions are mostly in the ventral temporal cortex (Fig. 4), e.g. the FFA, PPA, and LO. The existence of category-selective regions does not contradict with distributed category representation. Instead, the category specificity in a region is thought to emerge from its connectivity with other regions that also represent that category (Mahon and Caramazza, 2011), for processing domain-specific knowledge of particular importance to vision-guided action and cognition (Caramazza and Shelton, 1998).

Cortical category representations preserve semantic relationships

Our results suggest that cortical representational similarity between different categories is highly correlated with their semantic relationship (Fig. 6). In other words, the semantic relationship is preserved by cortical representation. This finding lends support for the notion of a continuous semantic space underlying the brain's category representation (Huth et al., 2012), which is a compelling hypothetical principle to bridge neural representation and linguistic taxonomy (Huth et al., 2016a). However, category information is not limited to semantic features, but includes hierarchically organized attributes that all define categories and their conceptual relationships. For example, "face" is not an isolated concept; it entails facial features ("eyes", "nose", "mouth"), each also having its own defining features. The similarity and distinction between categories may be attributable to one or multiple levels of features. In prior studies (Huth et al., 2012), the hierarchical nature of category information is not considered as every exemplar of each category is annotated by a pre-defined label. This causes an incomplete account of category representation, leaving it difficult to pinpoint which levels of category information drive the representational similarity between categories.

We have overcome this issue by extracting multiple layers of features from visual objects and evaluating the layer-wise contributions to cortical category representation. Our results show that similarity in cortical category representation is contributed by most layers of features, while different layers contributed differently to the representational modularity. Coarse categories (i.e. biological objects, non-biological objects, and background scenes) are most attributable to mid-level features, e.g. shapes, textures, and object parts (Fig. 7). In a finer level of categorization, terrestrial animals, aquatic animals, plants, and humans are most distinguishable in the semantic space; categorization of man-made and natural scenes is most supported by mid-level features (Fig. 9), likely reflecting the spatial layout of scene components (Kravitz et al., 2011; Harel et al., 2013).

Finer spatial scale supports more specific categorization

Our results suggest that object representations in the entire visual cortex support coarse categorization (Fig. 6), and representations in a smaller scale specific to each coarse category support subsequently finer categorization (Fig. 8). This finding is in line with the notion of nested spatial and representational hierarchies (Grill-Spector and Weiner, 2014): increasingly specific categorization results from category representations in a progressively finer spatial scale on the cortex. Such a spatial hierarchy describes a functional architecture that complies with both distributed (Haxby et al., 2001) and regional (Kanwisher et al., 1997; Peelen and Downing, 2005) representations of object knowledge, and their functional roles for categorization in multiple levels of specificity (Grill-Spector and Weiner, 2014). This nested hierarchy implies that widely distributed patterns of responses to visual objects are more distinguishable between coarsely defined categories than between relatively finer categories, as demonstrated in previous studies

(Kiani et al., 2007; Kriegeskorte et al., 2008). Finer object categorization may require representational differences in domain-specific regions (Kanwisher et al., 1997; Epstein and Kanwisher, 1998; Peelen and Downing, 2005).

The notion of spatial and representational hierarchies for graded categorization also has implications to decoding visual objects with multi-voxel pattern analysis (Haxby et al., 2001; Carlson et al., 2003; Cox and Savoy, 2003; Kriegeskorte et al., 2006; Pereira et al., 2009; Stansbury et al., 2013; Huth et al., 2016b). No single spatial scale is optimal for decoding visual objects across all levels of categories. The optimal spatial scale for decoding object categories depends on how specific the categories are defined.

Acknowledgement

The authors are thankful to Dr. Xiaohong Zhu and Dr. Byeong-Yeul Lee for constructive discussion, and Kuan Han for his assistance in collecting natural images. The research was supported by NIH R01MH104402 and Purdue University.

Reference

Adolf D, Weston S, Baecke S, Luchtmann M, Bernarding J, Kropf S (2014) Increasing the reliability of data analysis of functional magnetic resonance imaging by applying a new blockwise permutation method. Frontiers in neuroinformatics 8:72.

Afraz S-R, Kiani R, Esteky H (2006) Microstimulation of inferotemporal cortex influences face categorization. Nature 442:692-695.

Bracci S, de Beeck HO (2016) Dissociations and associations between shape and category representations in the two visual pathways. Journal of Neuroscience 36:432-444.

Cadieu CF, Hong H, Yamins DL, Pinto N, Ardila D, Solomon EA, Majaj NJ, DiCarlo JJ (2014) Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput Biol 10:e1003963.

Caramazza A, Shelton JR (1998) Domain-specific knowledge systems in the brain: The animate-inanimate distinction. Journal of cognitive neuroscience 10:1-34.

Carlson TA, Schrater P, He S (2003) Patterns of activity in the categorical representations of objects. Journal of cognitive neuroscience 15:704-717.

Chao LL, Martin A (2000) Representation of manipulable man-made objects in the dorsal stream. Neuroimage 12:478-484.

Chao LL, Haxby JV, Martin A (1999) Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. Nature neuroscience 2:913-919.

Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A (2016) Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Scientific reports 6.

Cox DD, Savoy RL (2003) Functional magnetic resonance imaging (fMRI)"brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage 19:261-270.

DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. Trends in cognitive sciences 11:333-341.

Eickenberg M, Gramfort A, Varoquaux G, Thirion B (2017) Seeing it all: Convolutional network layers map the function of the human visual system. NeuroImage 152:184-194.

Epstein R, Kanwisher N (1998) A cortical representation of the local visual environment. Nature 392:598-601.

Fellbaum C (1998) WordNet: Wiley Online Library.

Fox CJ, Iaria G, Barton JJ (2009) Defining the face processing network: optimization of the functional localizer in fMRI. Human brain mapping 30:1637-1651.

Freud E, Plaut DC, Behrmann M (2016) 'What'Is Happening in the Dorsal Visual Pathway. Trends in Cognitive Sciences 20:773-784.

Gallese V, Lakoff G (2005) The brain's concepts: The role of the sensory-motor system in conceptual knowledge. Cognitive neuropsychology 22:455-479.

Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, Xu J, Jbabdi S, Webster M, Polimeni JR (2013) The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage 80:105-124.

Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF, Jenkinson M (2016) A multi-modal parcellation of human cerebral cortex. Nature.

Gómez S, Jensen P, Arenas A (2009) Analysis of community structure in networks of correlated data. Physical Review E 80:016114.

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

Güçlü U, van Gerven MA (2015a) Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. NeuroImage.

Güçlü U, van Gerven MA (2015b) Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. Journal of Neuroscience 35:10005-10014.

Harel A, Kravitz DJ, Baker CI (2013) Deconstructing visual scenes in cortex: gradients of object and spatial layout information. Cerebral Cortex 23:947-957.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425-2430.

Haxby JV, Guntupalli JS, Connolly AC, Halchenko YO, Conroy BR, Gobbini MI, Hanke M, Ramadge PJ (2011) A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72:404-416.

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770-778.

Horikawa T, Kamitani Y (2017) Generic decoding of seen and imagined objects using hierarchical visual features. 8:15037.

Huth AG, Nishimoto S, Vu AT, Gallant JL (2012) A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76:1210-1224.

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016a) Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532:453-458.

Huth AG, Lee T, Nishimoto S, Bilenko NY, Vu AT, Gallant JL (2016b) Decoding the semantic content of natural movies from human brain activity. Frontiers in systems neuroscience 10.

Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. Journal of neuroscience 17:4302-4311.

Kay KN, Naselaris T, Prenger RJ, Gallant JL (2008) Identifying natural images from human brain activity. Nature 452:352-355.

Khaligh-Razavi S-M, Kriegeskorte N (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol 10:e1003915.

Kiani R, Esteky H, Mirpour K, Tanaka K (2007) Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. Journal of neurophysiology 97:4296-4309.

Kourtzi Z, Connor CE (2011) Neural representations for object perception: structure, category, and adaptive coding. Annual review of neuroscience 34:45-67.

Kravitz DJ, Peng CS, Baker CI (2011) Real-world scene representations in high-level visual cortex: it's the spaces more than the places. Journal of Neuroscience 31:7322-7333.

Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. Proceedings of the National academy of Sciences of the United States of America 103:3863-3868.

Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60:1126-1141.

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097-1105.

Leacock C, Chodorow M (1998) Combining local context and WordNet similarity for word sense identification. WordNet: An electronic lexical database 49:265-283.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436-444.

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, pp 281-297: Oakland, CA, USA.

Mahon BZ, Caramazza A (2011) What drives the organization of object knowledge in the brain? Trends in cognitive sciences 15:97-103.

Mahon BZ, Anzellotti S, Schwarzbach J, Zampini M, Caramazza A (2009) Category-specific organization in the human brain does not require visual experience. Neuron 63:397-405.

Martin A (2007) The representation of object concepts in the brain. Annu Rev Psychol 58:25-45.

Mur M, Ruff DA, Bodurka J, De Weerd P, Bandettini PA, Kriegeskorte N (2012) Categorical, yet graded–single-image activation profiles of human category-selective cortical regions. Journal of Neuroscience 32:8649-8662.

Naselaris T, Stansbury DE, Gallant JL (2012) Cortical representation of animate and inanimate objects in complex natural scenes. Journal of Physiology-Paris 106:239-249.

Naselaris T, Kay KN, Nishimoto S, Gallant JL (2011) Encoding and decoding in fMRI. Neuroimage 56:400-410.

Naselaris T, Olman CA, Stansbury DE, Ugurbil K, Gallant JL (2015) A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. Neuroimage 105:215-228.

Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL (2011) Reconstructing visual experiences from brain activity evoked by natural movies. Current Biology 21:1641-1646.

Peelen MV, Downing PE (2005) Selectivity for the human body in the fusiform gyrus. Journal of neurophysiology 93:603-608.

Pereira F, Mitchell T, Botvinick M (2009) Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45:S199-S209.

Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I (2005) Invariant visual representation by single neurons in the human brain. Nature 435:1102-1107.

Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nature neuroscience 2:1019-1025.

Smith SM, Beckmann CF, Andersson J, Auerbach EJ, Bijsterbosch J, Douaud G, Duff E, Feinberg DA, Griffanti L, Harms MP (2013) Resting-state fMRI in the human connectome project. Neuroimage 80:144-168.

Stansbury DE, Naselaris T, Gallant JL (2013) Natural scene statistics account for the representation of scene categories in human visual cortex. Neuron 79:1025-1034.

Van Essen DC, Anderson CH, Felleman DJ (1992) Information processing in the primate visual system: an integrated systems perspective. Science 255:419.

Wen H, Shi J, Zhang Y, Lu K-H, Liu Z (2016) Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. arXiv preprint arXiv:160803425.

Yamins DL, DiCarlo JJ (2016) Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience 19:356-365.

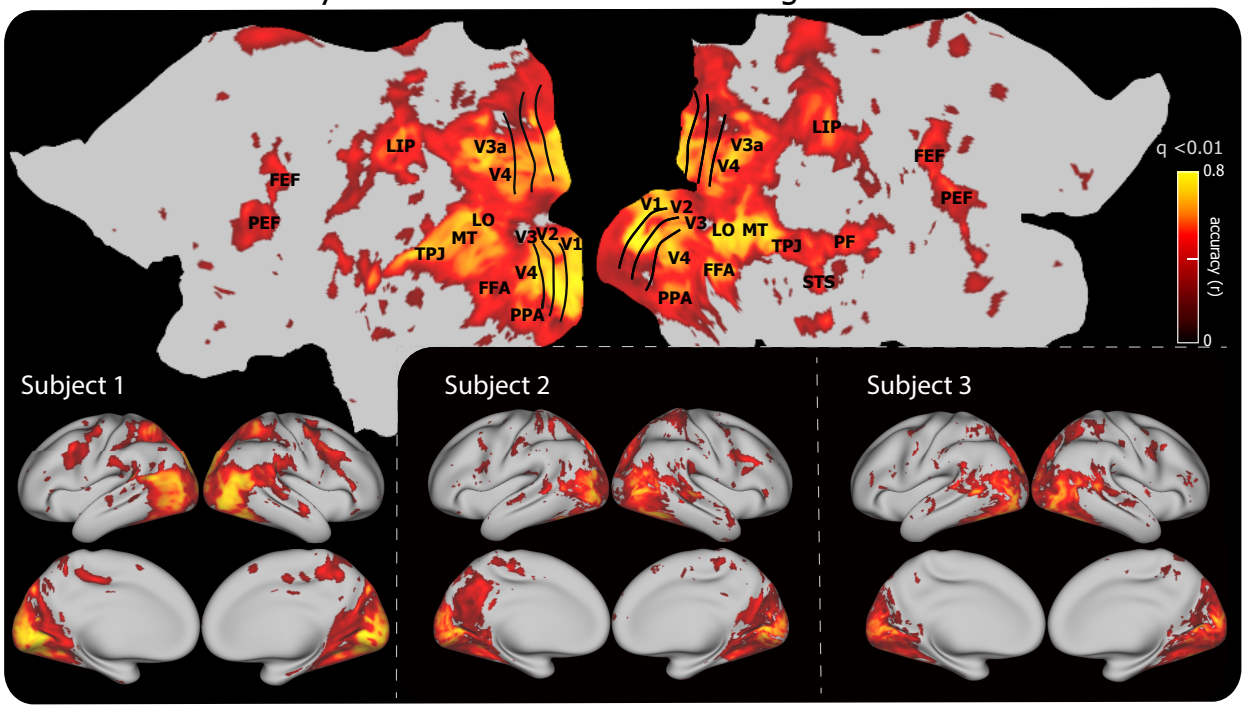
Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences 111:8619-8624.

Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H (2015) Understanding neural networks through deep visualization. arXiv preprint arXiv:150606579.

Table 1 Visual object categories

airplane	bag	ball	beach	bear	bedroom	bike	bird
boat	book	bottle	bowl	bridge	building	car	cat
cellphone	chair	chicken	classroom	computer	corridor	sports court	cup
dog	dolphin	door	drink	elephant	factory	falls	fish
flag	flower	forest	fruit	goose	grass	hat	horse
house	instrument	kitchen	knife	lion	living room	man	market
microphone	monkey	mountain	office	restaurant	toilet	river	road
stone	shark	sheep	ship	shoe	sign	sky	snake
snow	street	sunrise	table	tiger	tools	stoplight	tree
turtle	umbrella	vegetable	TV	watch	wave	woman	vending machine

a. Prediction accuracy of the ResNet-based encoding models



b. Comparison in the prediction accuracy between ResNet and AlexNet

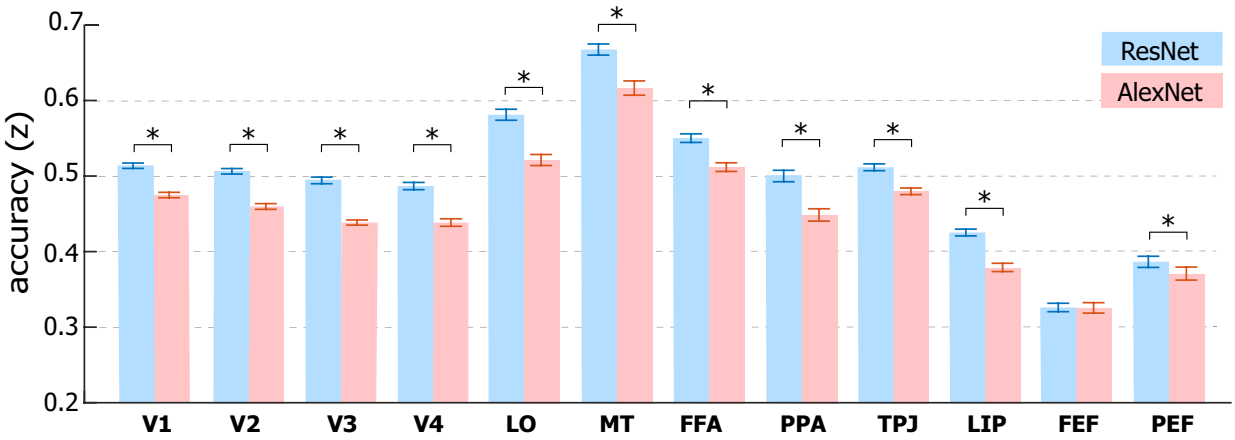


Figure1. DNN-based Voxel-wise encoding models. (a) Performance of ResNet-based encoding models in predicting the cortical responses to novel testing movies for three subjects. The accuracy is measured by the average Pearson's correlation coefficient (r) between the predicted and the observed fMRI responses across five testing movies ($q<0.01$ after correction for multiple testing using the false discovery rate (FDR) method). The prediction accuracy is displayed on both flat (top) and inflated (bottom left) cortical surfaces for Subject 1. **(b)** Comparison between

the ResNet-based and the AlexNet-based encoding models. Each bar represents the mean±SE of the prediction accuracy (Fisher's z-transformation of r) within a ROI across voxels and subjects, and * represents a significance p-value ($p<0.001$) with paired t-test.

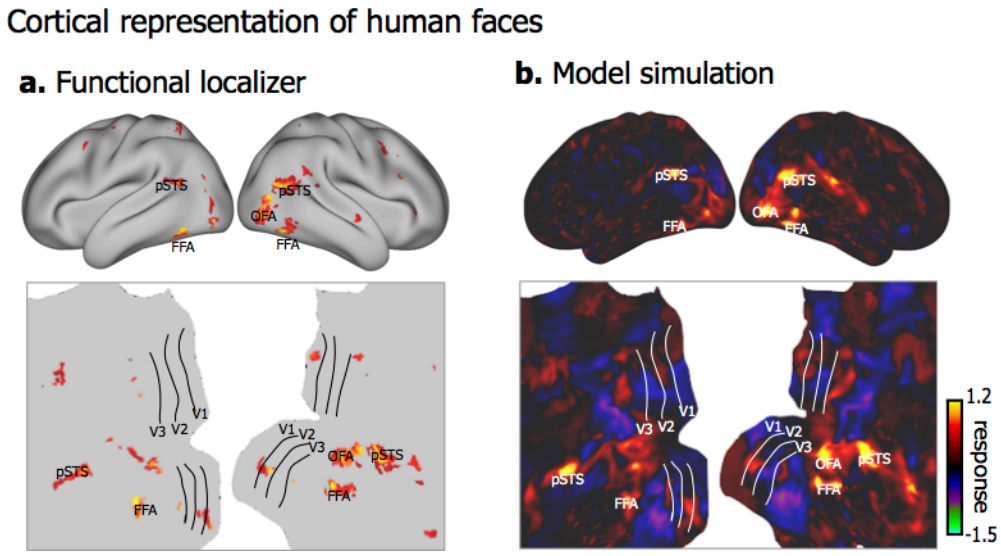


Figure2. Human-face representations with encoding models and functional localizer (a) Localizer activation maps comprising regions selective for human faces, including occipital face area (OFA), fusiform face area (FFA), and posterior superior temporal sulcus (pSTS). **(b)** Model-simulated representation of human face from ResNet-based encoding models. The representation is displayed on both inflated (top) and flat (bottom) cortical surfaces.

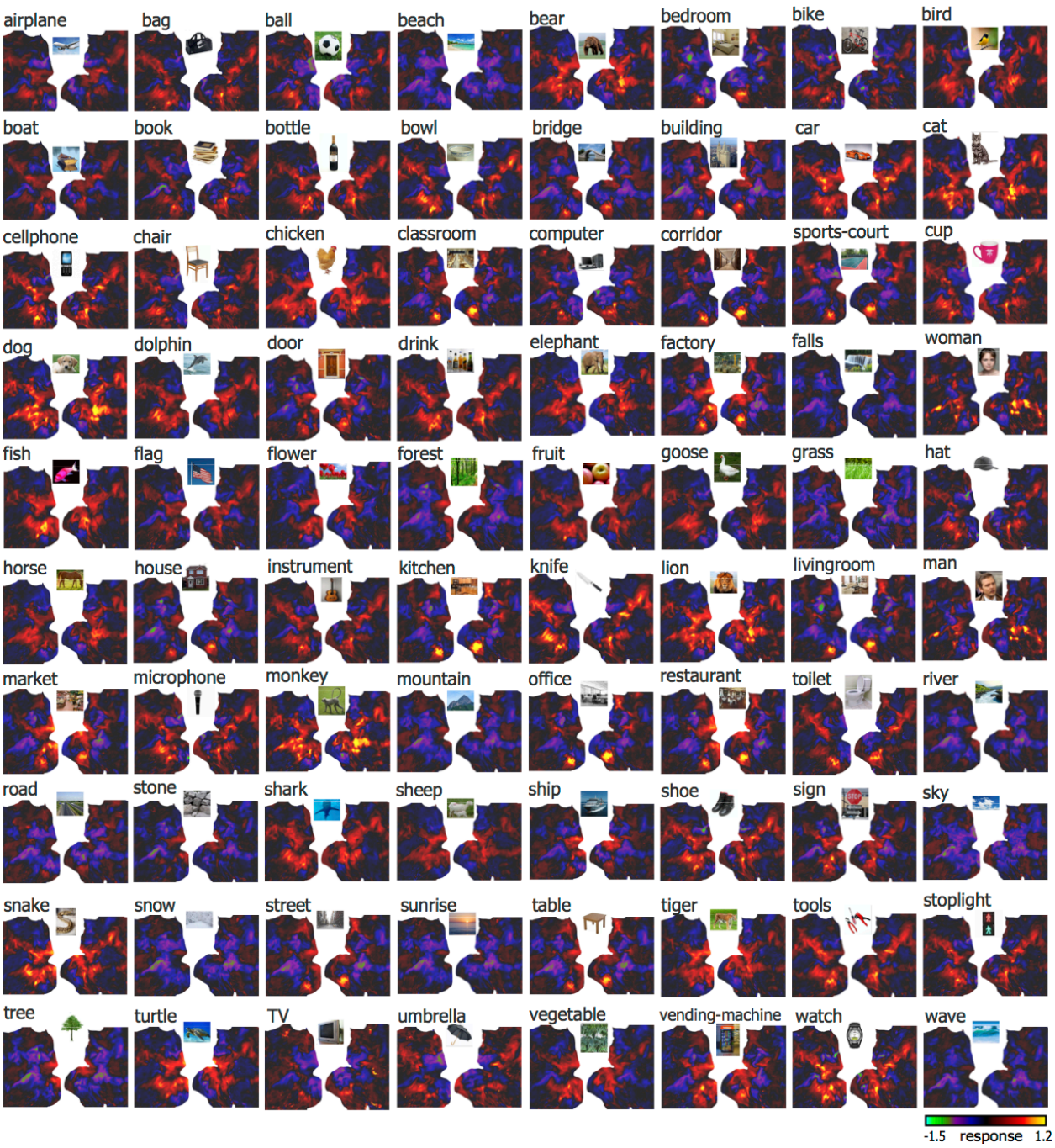


Figure3. Cortical representations of 80 object categories. Each panel shows the representation map of an object category on flat cortical surface from Subject 1. The category label is on top left, and the image on the top middle is an exemplar of the category. The color bar shows the cortical response.

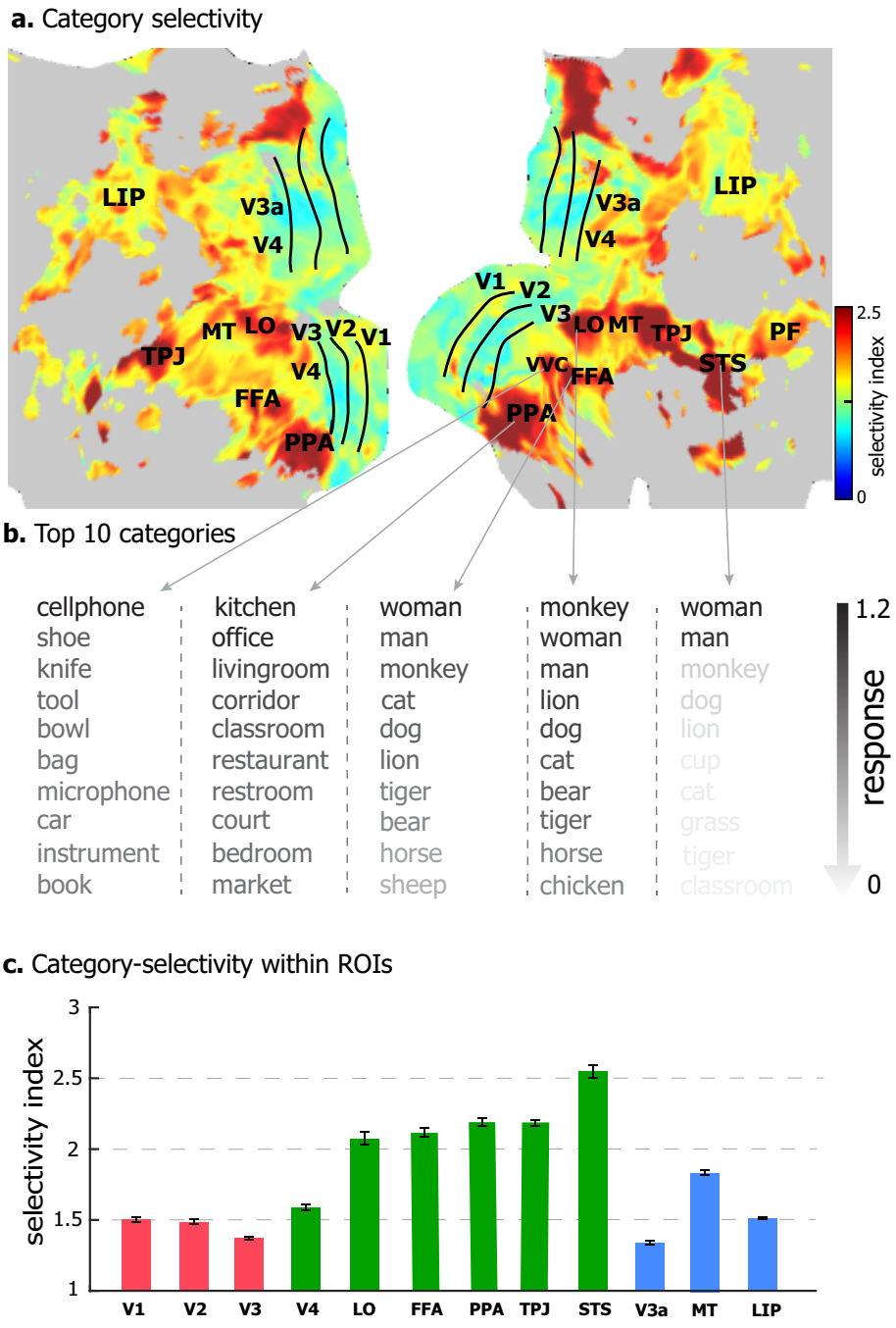


Figure4. Category-selectivity at individual cortical locations. (a) The category-selectivity across the cortical surface. **(b)** The category-selectivity profile of example cortical locations. For each location, top 10 categories with the highest responses are showed in descending order. **(c)** Category-selectivity within ROIs (mean±SE) in the early visual areas (red), ventral stream areas (green), and dorsal stream areas (blue).

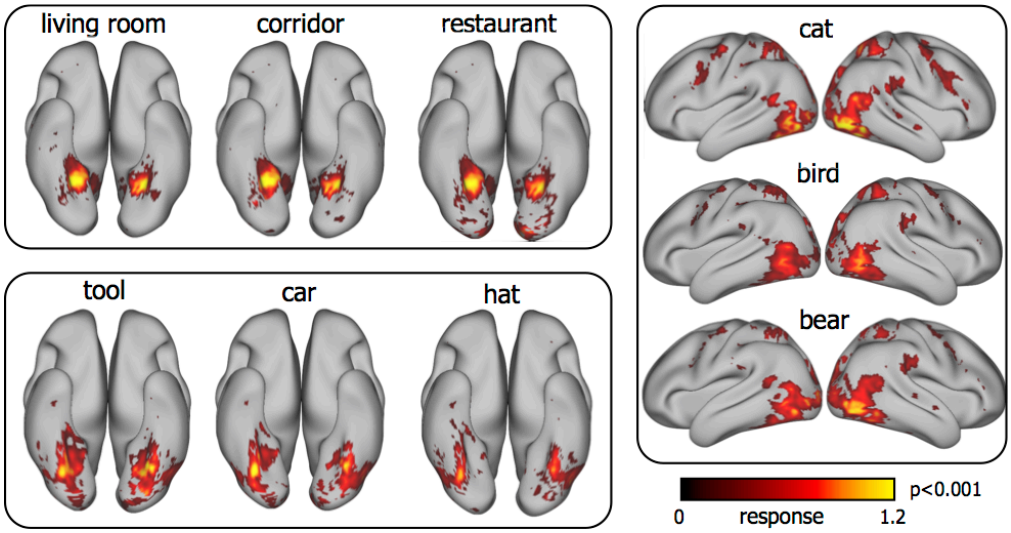


Figure5. Example categories that shared similar cortical representations. The cortical representations are displayed in best view on the inflated cortical surface. It was thresholded by assessing the significance of the response to a category against 50,000 random and non-selective natural pictures with two-sample t-test ($p<0.001$, Bonferroni correction for the number of voxels).

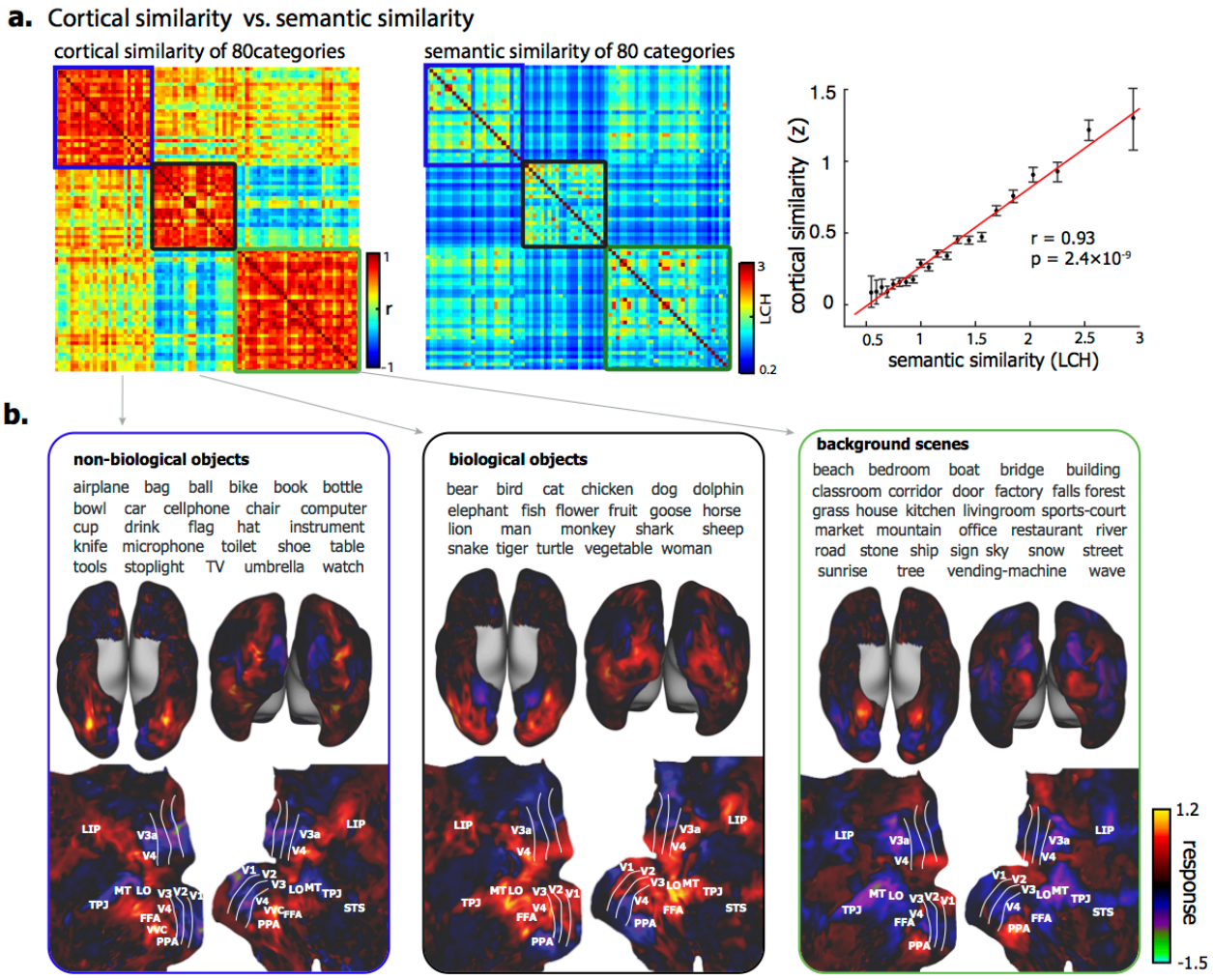
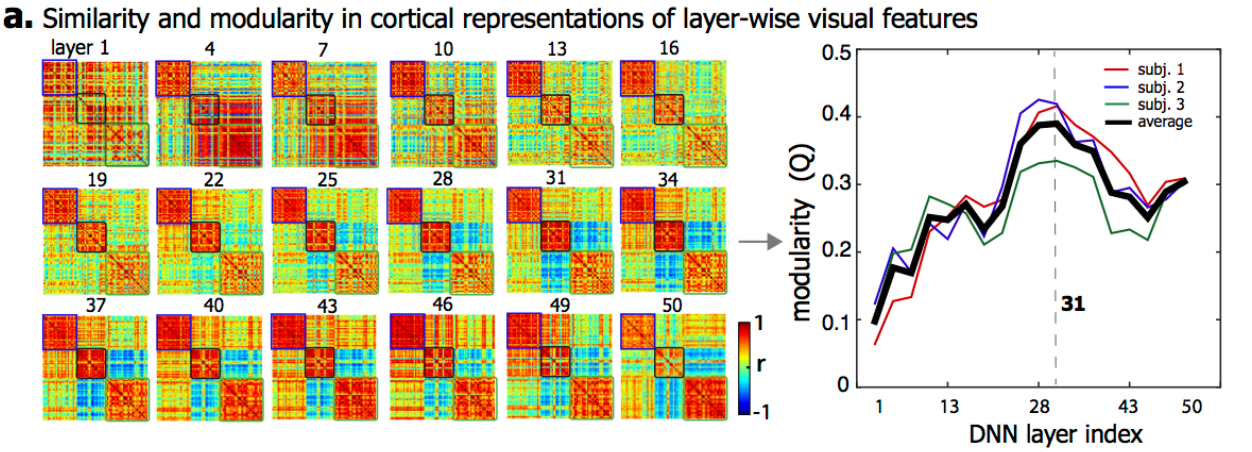


Figure6. Categorical similarity and modularity in cortical representation at the scale of the entire visual cortex. (a) The left is the similarity matrix (r) of the cortical representations between categories. Each element represents the average cortical similarity between a pair of categories across subjects. It is well separated into three modules with modularity Q=0.35. The middle is the similarity matrix of the semantic content between categories (measured by LCH). The right is a plot of the mean±SE of cortical similarity (Fisher's z-transformation of r) vs. the semantic similarity (LCH takes discrete values). **(b)** These three modules are related to three superordinate-level categories: non-biological objects, biological objects, and background

scenes. The average cortical representations across categories within modules are showed in the bottom on both inflated and flat cortical surfaces.



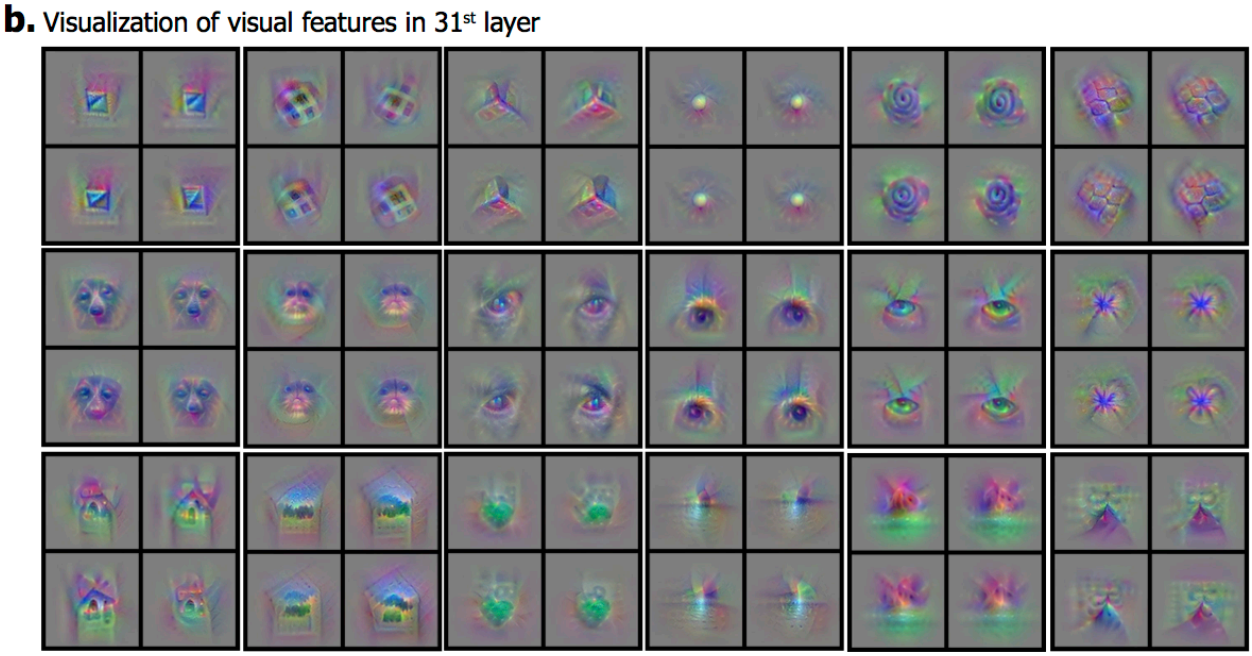


Figure7. Contribution of layer-wise visual features to the similarity and modularity in cortical representation. (a) The left shows the similarity between categories in the cortical representations that are contributed by separated category information from individual layers. The order of categories is the same as in Figure 6.a. The right plot shows the modularity index across all layers. The visual features at the middle layers have the highest modularity. **(b)** 18 example visual features at the 31st layer are visualized in pixel space. Each visual feature shows 4 exemplars that maximize the feature representation (see ***Layer-wise contribution to cortical categorical representation*** in **Materials and Methods**).

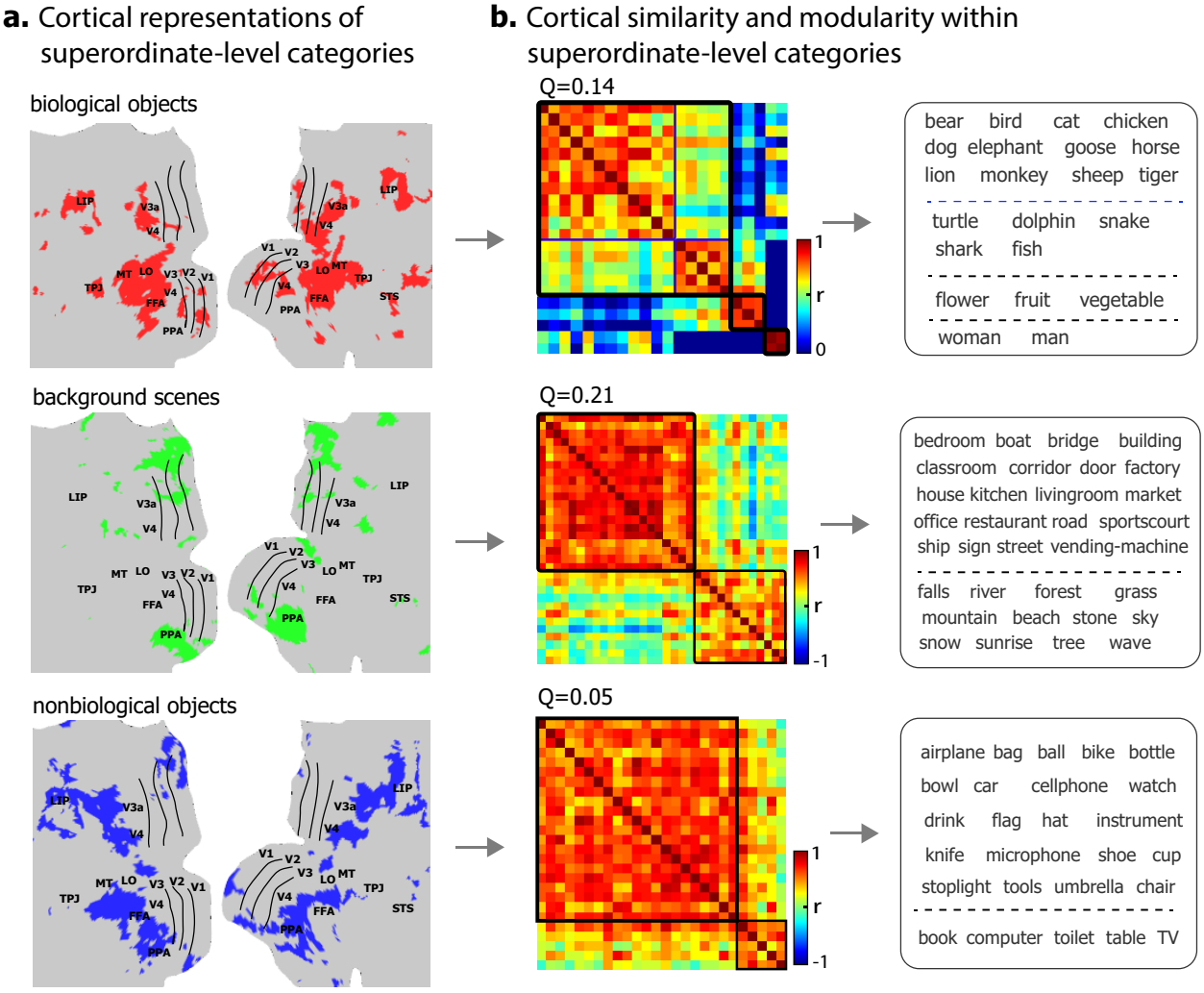
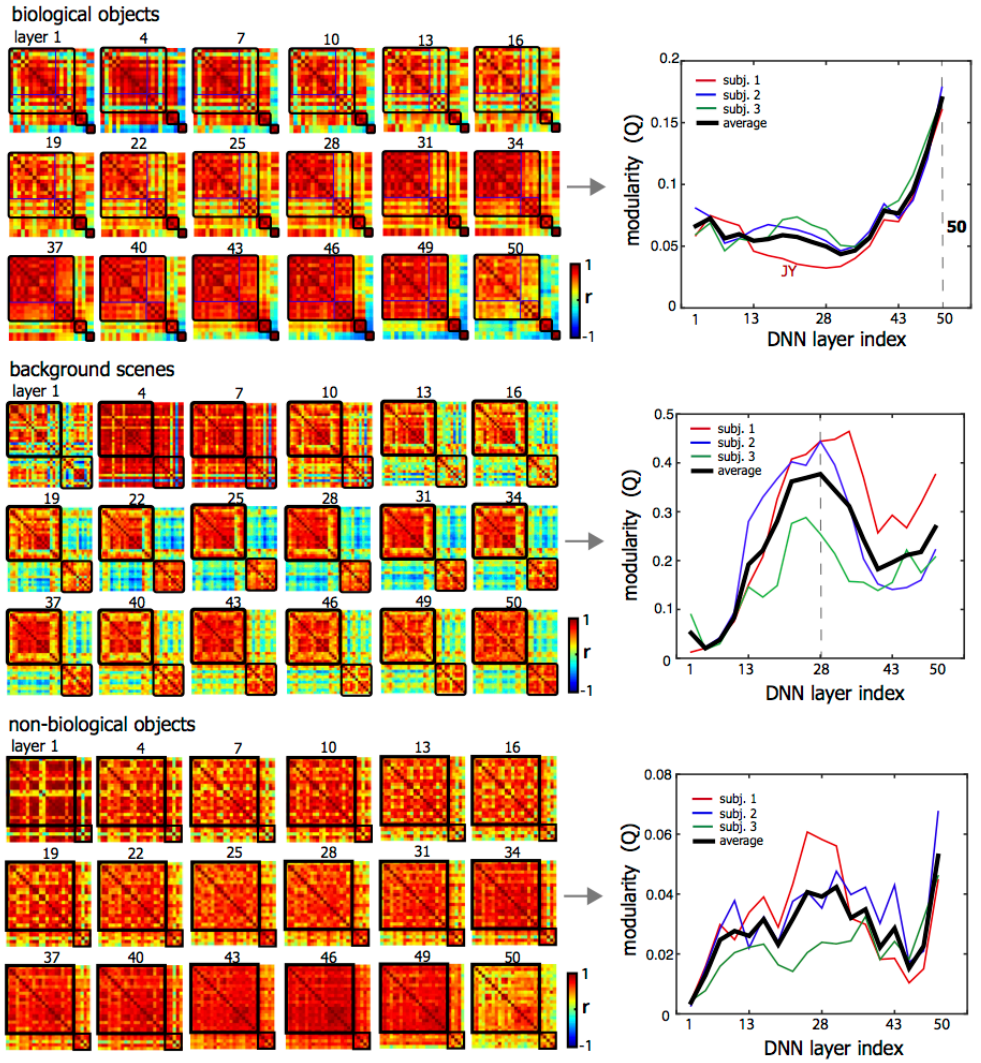


Figure8. Categorical similarity and modularity in cortical representation within superordinate-level categories. (a) Fine-scale cortical areas specific to each superordinate-level category: biological objects (red), background scenes (green) and non-biological objects (blue). **(b)** The cortical similarity between categories in fine-scale cortical representation. The categories in each sub-module were displayed on the right. See Figure 10 for individual subjects.

a. Similarity and modularity in cortical representations of layer-wise visual features



b. Visualization of features in 50th layer

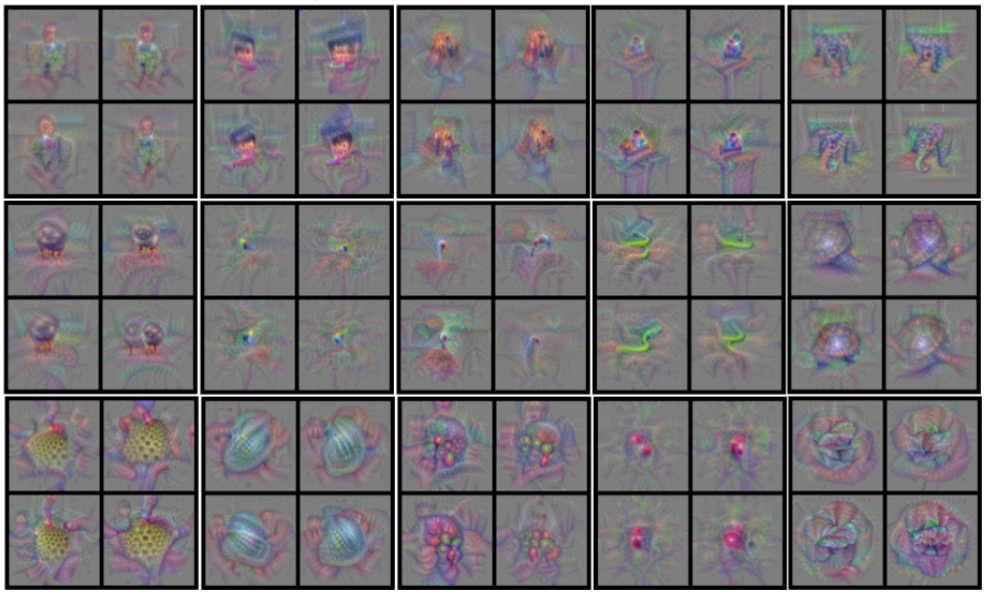


Figure9. Contribution of layer-wise visual features to the similarity and modularity in cortical representations within superordinate-level categories. (a) The left shows the similarity between categories in fine-scale cortical representations that are contributed by separated category information from individual layers. The order of categories is the same as in Figure 8. The right plot shows the modularity index across all layers. The highest-layer visual features show the highest modularity for biological objects. **(b)** 15 example visual features at the 50^{st} layer are visualized in pixel space. Each visual feature showed 4 exemplars that maximize the feature representation (see ***Layer-wise contribution to cortical categorical representation*** in **Materials and Methods**).

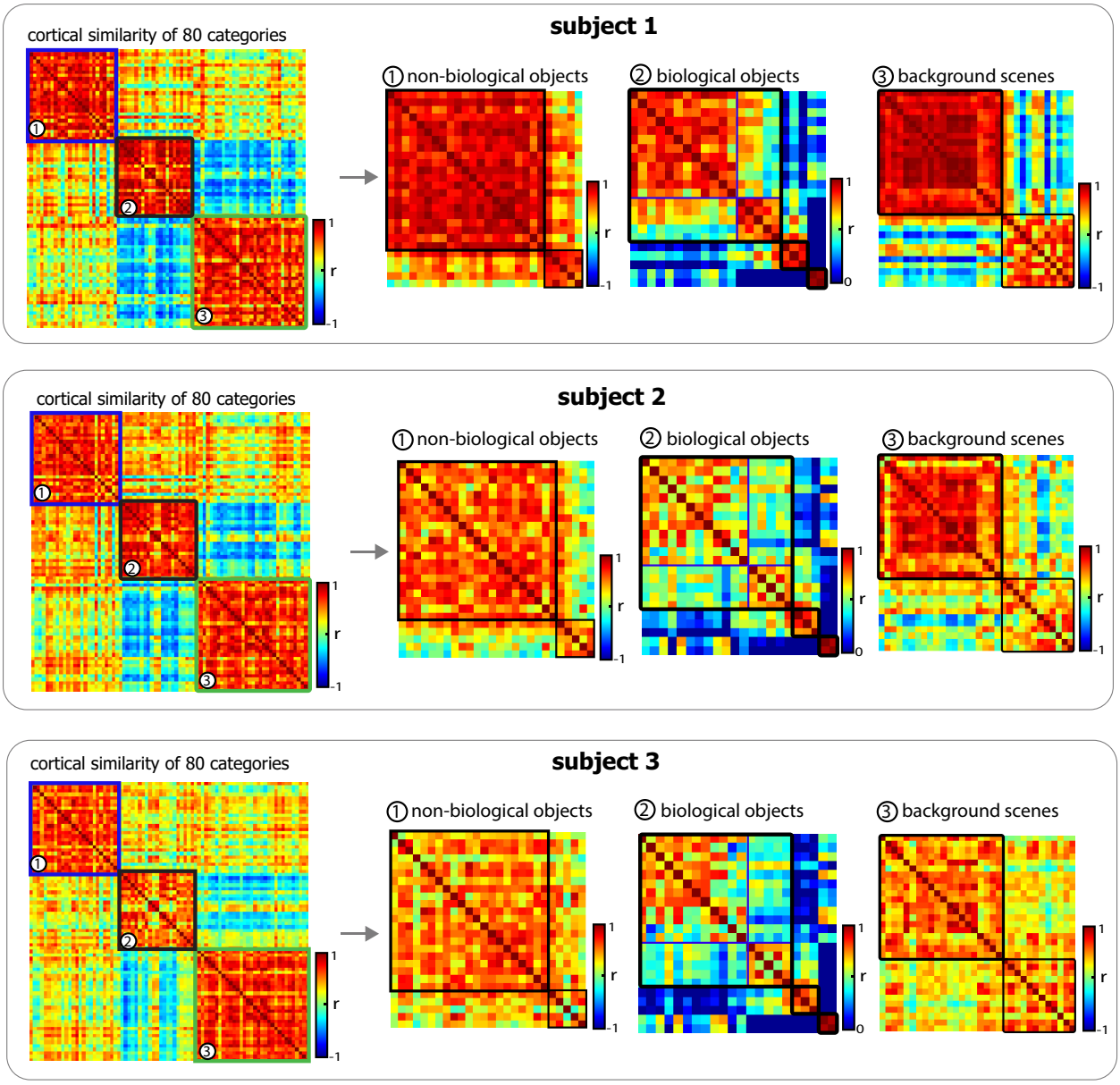


Figure10. Categorical hierarchy for individual subjects. For each subject, the left shows the similarity between categories in the cortical representation at the scale of the entire visual cortex, and the order of categories is the same as in Figure 6.a. The right shows the similarity within each superordinate-level category in the finer-scale cortical representations, and the orders of categories are the same as in Figure 8.