NeuroImage 197 (2019) 368–382

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images

Marcie L. King [a,b,1], Iris I.A. Groen [a,c,1], Adam Steel [a], Dwight J. Kravitz [d], Chris I. Baker [a,*]

[a] Laboratory of Brain and Cognition, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, 20892, USA
[b] Department of Psychological and Brain Sciences, University of Iowa, W311 Seashore Hall, Iowa City, IA, 52242, USA
[c] Department of Psychology, New York University, 6 Washington Place, New York, NY, 10003, USA
[d] Department of Psychology, George Washington University, 2125 G St. NW, Washington, DC, 20008, USA

A B S T R A C T

Numerous factors have been reported to underlie the representation of complex images in high-level human visual cortex, including categories (e.g. faces, objects, scenes), animacy, and real-world size, but the extent to which this organization reflects behavioral judgments of real-world stimuli is unclear. Here, we compared representations derived from explicit behavioral similarity judgments and ultra-high field (7T) fMRI of human visual cortex for multiple exemplars of a diverse set of naturalistic images from 48 object and scene categories. While there was a significant correlation between similarity judgments and fMRI responses, there were striking differences between the two representational spaces. Behavioral judgements primarily revealed a coarse division between man-made (including humans) and natural (including animals) images, with clear groupings of conceptually-related categories (e.g. transportation, animals), while these conceptual groupings were largely absent in the fMRI representations. Instead, fMRI responses primarily seemed to reflect a separation of both human and non-human faces/bodies from all other categories. Further, comparison of the behavioral and fMRI representational spaces with those derived from the layers of a deep neural network (DNN) showed a strong correspondence with behavior in the top-most layer and with fMRI in the mid-level layers. These results suggest a complex relationship between localized responses in high-level visual cortex and behavioral similarity judgments – each domain reflects different properties of the images, and responses in high-level visual cortex may correspond to intermediate stages of processing between basic visual features and the conceptual categories that dominate the behavioral response.

1. Introduction

The ventral visual pathway, extending from primary visual cortex (V1) down the temporal lobe into ventral temporal cortex, is thought to be critical for object, face and scene recognition (Kravitz et al., 2013). While posterior regions in this pathway respond strongly to the presentation of low-level visual features, more anterior regions are thought to encode high-level categorical aspects of the visual input. For example, functional magnetic resonance imaging (fMRI) studies have identified category-selective regions in ventral temporal cortex (vTC) and lateral occipitotemporal cortex (lOTC) that show preferential responses for images of one category compared to another (e.g. face-selective fusiform face area or FFA, scene-selective parahippocampal place area or PPA, and object-selective lateral occipital complex or LOC; (Kanwisher and Dilks, 2013). However, many other factors have been reported to contribute to responses in high-level visual cortex, including, but not limited to, eccentricity (Hasson et al., 2003), elevation (Silson et al., 2015), real-world size (Konkle and Oliva, 2012), typicality (Iordan et al., 2016), category level (i.e. superordinate, basic, subordinate – (Iordan et al., 2015)), complexity (Güçlütürk et al., 2018), semantic similarity (Carlson et al., 2014), and animacy (Kriegeskorte et al., 2008; Connolly et al., 2012; Naselaris et al., 2012; Sha et al., 2015; Proklova et al., 2016) with a shared representational space across participants (Haxby et al., 2011; Guntupalli et al., 2016; Van Uden et al., 2018). The goal of the current study was to assess the correspondence between the fMRI response of high-level visual cortex and mental representations of category elicited by naturalistic visual stimuli by comparing the representational space reflected in fMRI responses with behavioral similarity judgements across a broad range of object and scene categories.

Determining how responses in high-level visual cortex relate to behavioral judgments is critical for elucidating the functional significance of these regions. The presence of responses in both human and non-human primate visual cortex specific to a given category or type of object suggest there may be a close correspondence between responses in high-level visual cortex and behavioral similarity judgments (Kravitz et al., 2013; Grill-Spector and Weiner, 2014). However, there are many behavioral goals beyond identification that these regions are likely to support and many possible dimensions on which visual input can be

* Corresponding author.
E-mail address: bakerchris@mail.nih.gov (C.I. Baker).
[1] co-first authors.

https://doi.org/10.1016/j.neuroimage.2019.04.079
Received 22 October 2018; Received in revised form 26 March 2019; Accepted 29 April 2019
Available online 1 May 2019
1053-8119/Published by Elsevier Inc.

evaluated (Malcolm et al., 2016; Peelen and Downing, 2017). Further it has been suggested that behavioral and fMRI measurements provide complementary insights into conceptual organization with fMRI measures revealing intrinsic properties of individual concepts and behavioral measures the overall structure of the knowledge for a domain of concepts (Bauer and Just, 2019). While the fMRI responses in both human and non-human primate vTC appear to reflect major distinctions between animate/inanimate and face/body, behavioral similarity judgements in a free arrangement task reveal additional fine-grained representational structure, particularly for inanimate objects (Kriegeskorte et al., 2008; Mur et al., 2013). However, these studies contained a limited sampling of different categories that emphasized some categories (e.g. faces, food/-fruit) over others (e.g. chairs, appliances) and may have only captured part of the representational space. Thus, it is unclear to what extent vTC reflects the broader representational space across more diverse categories. While other fMRI studies have included a broader sampling of different categories (Huth et al., 2012; Naselaris et al., 2012), behavioral judgments were not collected beyond labels for discrete elements of the images that may not adequately characterize the nature of conceptual representations. Here, we combined a diverse sampling of different categories with both ultra-high field (7T) fMRI and detailed behavioral similarity measurements to determine what aspects of representation are shared between behavior and the responses of high-level visual cortex.

We presented multiple naturalistic images from 48 categories ranging across both object (e.g. bags, dolls) and scene (e.g. kitchens, mountains) categories. In contrast to some prior studies that presented segmented objects with limited, arbitrary or no context (Kriegeskorte et al., 2008; Konkle and Oliva, 2012) (Yamins et al., 2014) our study presented objects in typical contexts (Cichy et al., 2019). For behavior, we adopted the same paradigm used in other studies (Kriegeskorte et al., 2008; Kriegeskorte and Mur, 2012; Mur et al., 2013; Cichy et al., 2019) in which participants arranged the stimuli in a two dimensional space according to similarity, without being explicitly told to focus on specific dimensions. This task allows us to assess participants' implicit understanding of the stimuli, reflecting the integration of multiple dimensions for each image for comparison with all the other images across the stimulus set. Consistent with prior studies we found highly reproducible representations in both behavior and fMRI that were significantly correlated. However, there was little evidence for the previously reported animacy division and close inspection of the representational spaces captured by similarity judgments and fMRI revealed that their overall geometry was in fact strikingly different. In particular, behavioral judgments reflected a manmade/natural division with groupings of conceptually related categories, while cortical regions largely showed a separation of images containing human and non-human faces and bodies from everything else. Computational features extracted from a deep neural network (DNN) trained on object recognition correlated with representational structure in both behavior and fMRI, but the strongest match with behavior was with the highest DNN layer, while fMRI correlated best with a mid-level DNN layer. Collectively, these results show that while there is a significant correlation between the behavioral and fMRI data from high-level visual cortex, there is a striking difference in the structure of the representational space with their correspondence limited to a small number of categories.

2. Materials and Methods

Stimuli. We retrieved high-resolution (1024x768 pixels) color photographs from Google Images to construct two sets of stimuli, each comprised of 144 individual color images of complex scenes. We included two separate sets to test generalization of our findings across images. Each set of images (hereby referred to as Image Set 1 and Image Set 2) contained 48 concrete categories, with 3 exemplar images per category. We first chose 48 category labels to reflect a diverse range of common, naturalistic object and scene categories, including category labels that had been used in prior studies. Then for each label we selected six images that depicted that label. All of the images in Image Set 1 and Image Set 2 depicted people, places, and things in natural context and from familiar viewpoints. The images portrayed scenes that one might expect to see on a typical day, and were chosen for their neutral nature (i.e. to be unlikely to elicit any strong emotional response). Fig. 1 shows a single exemplar from each of the 48 categories. The full set of images used in the study are shown in Supplementary Figs. 1 and 2.

Participants and testing. Twenty healthy human volunteers (9 male, mean age = 27.7 years) participated in the behavioral similarity judgment experiment. Ten participants viewed Image Set 1 (4 male, mean age = 29.3) and 10 participants viewed Image Set 2 (5 male, mean age = 26.1). Ten of these participants also participated in the corresponding fMRI experiment prior to participating in the behavioral portion of this study. Five of these participants viewed stimuli from Image Set 1 (3 male, mean age = 26.6 years) and five participants viewed stimuli from Image Set 2 (2 male, mean age = 26.2 years). Each participant saw the same stimulus set in both the behavioral and fMRI experiment. All fMRI participants completed the fMRI scan session before rating the behavioral similarities of the images. Each scan session lasted 2 h and the behavioral similarity experiment lasted 1 h. This study was conducted in accordance with The National Institutes of Health Institutional Review Board, and all participants gave written informed consent as part of the study protocol (93 M-0170, NCT00001360) prior to participation in the study.

Behavioral paradigm. We adopted a multi-arrangement paradigm previously used by Kriegeskorte, Mur and colleagues (Kriegeskorte and Mur, 2012; Mur et al., 2013). Participants were seated at a distance of approximately 50 cm in front of a computer monitor (Dell U3014, 30 inches, 2560 × 1600 pixels) and completed the object arrangement task on 144 images comprising either Image Set 1 or Image Set 2. At the onset of the task, all 144 images were presented simultaneously in random order around the perimeter of a circle presented on the computer monitor, forming an "arena" in which similarity judgments were made. Participants were instructed to "please arrange these images according to their similarity, whatever that means to you. Images that are more similar should go closer together and images that are less similar should go farther apart." These instructions were purposefully general so as not to bias the arrangements of the images in any particular way, allowing us to investigate what dimensions participants spontaneously use when judging the similarity between images. Thus, we assume that each participant's arrangement reflects the integration of multiple dimensions that vary across the stimulus set. Participants dragged the individual images into the arena using the mouse and physically arranged them according to their perceived similarity. Given the large number of images (and thus the small size each could be presented at), when a participant clicked on a particular image in the arena, an enlarged version of the image (150 × 200 pixels) was displayed in the top right of the computer screen.

Given the large number of images in the stimulus sets, participants completed only one arrangement of the images, in contrast to the original implementation of this method that used additional trials with selective subsets of stimuli (Kriegeskorte and Mur, 2012). In our experience we have found very high correlations between results on the first and last trial when running multiple trials ($r > 0.9$ for a set of 48 objects with 83 trials). Further, participants were able to re-arrange images within the circular area on the screen after their initial placement as many times as they wanted within a 1-h time limit, and they were encouraged to verify that they were satisfied with the final arrangement. One of the benefits of this arrangement method is that we were able to collect a large number of simultaneous pairwise similarity judgments in a reasonably short amount of time within individual participants, allowing comparison of fMRI and similarity data at the individual participant level. Perceived object-similarity is traditionally measured using pairwise similarity judgments, however it would take many hours and testing sessions to acquire judgments on our 10,296 possible pair combinations of images in an individual participant. Therefore, in the current method we used the

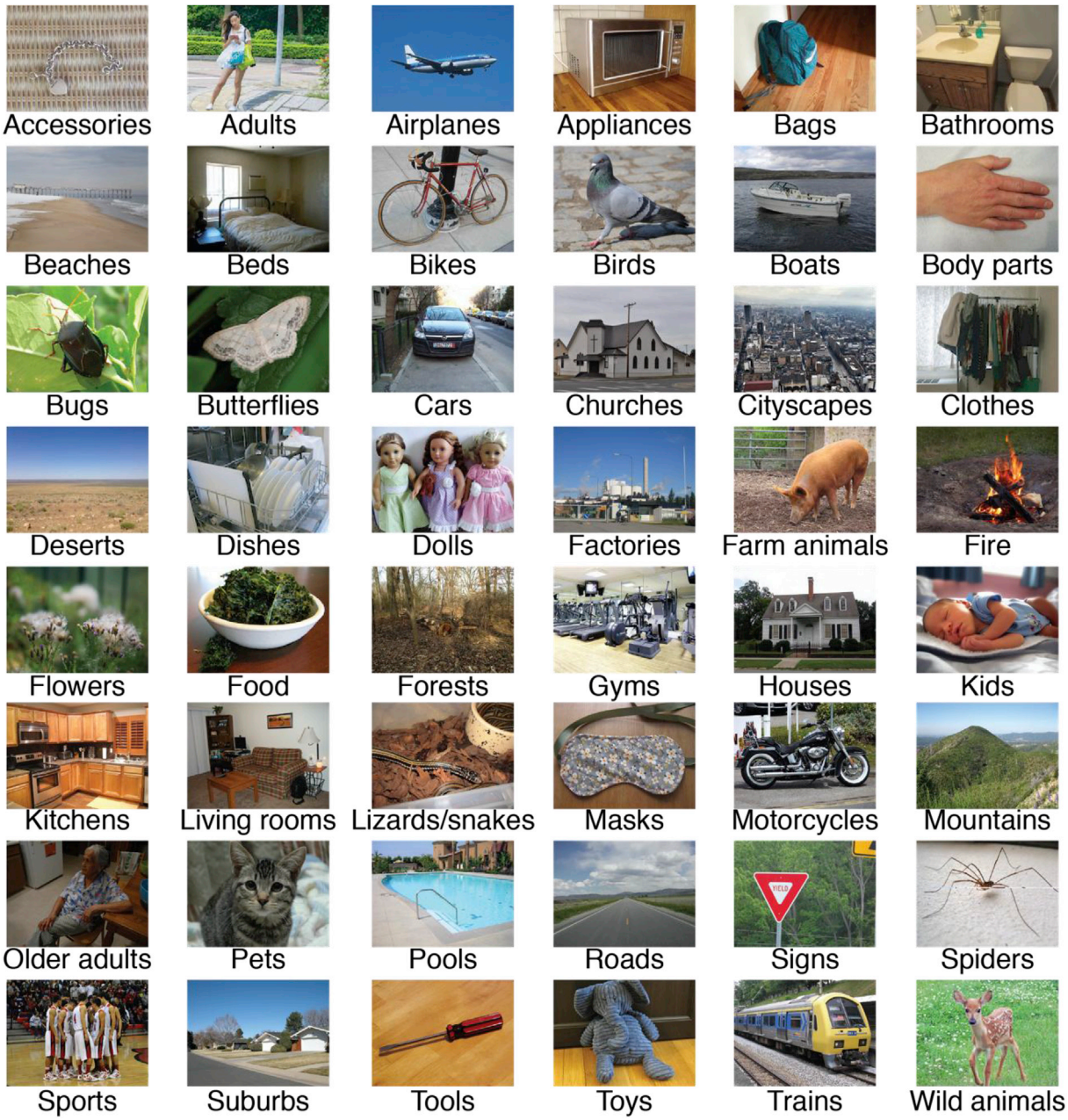

Fig. 1. Naturalistic image categories. One exemplar from each of the 48 image categories, presented in alphabetical order.

spatial arrangement of the images as a measure of their perceived similarity. Specifically, the Euclidean distance between an image and every other image was used as the measurement of perceived dissimilarity between the images (i.e. dissimilarity estimate). Representational dissimilarity matrices (RDMs) were constructed for each participant, using the ranked dissimilarity estimates for each image pair. Note that the distance matrix discards the absolute position of stimuli and only retains their relative location, which should minimize bias related to the initial placement of the stimuli.

fMRI paradigm. Participants were scanned while viewing the stimuli on a back-projected screen through a rear-view mirror that was mounted on the head coil. Stimuli were presented at a resolution of 1024×768 pixels and subtended 20×15 degrees of visual angle. Individual scenes were presented in an event-related design for a duration of 500 ms, separated by a 5 s interval. Throughout the experimental run, a small fixation cross ($<0.5^{\circ}$) was presented in the center of the screen. Participants viewed all 144 images in either Image Set 1 or Image Set 2 while performing an unrelated fixation cross task. We chose an unrelated task rather than having participants perform a task on the images themselves to minimize any interaction between the task and the images. While this fixation task draws some attention away from the images, the sudden onset large, colorful images automatically capture attention and participants could not exclusively focus on the fixation task at the expense of the images. Prior studies investigating the representational space of object and scenes with fMRI have also used similar fixation tasks (Kriegeskorte et al., 2008; Kravitz et al., 2011). For this task, simultaneous with the onset of each stimulus, either the vertical or horizontal arm of the fixation cross became slightly elongated. Participants were asked to indicate, via button response, whether the horizontal or vertical line of the fixation cross was longer. Both arms changed equally often within a given run, and arm changes were randomly assigned to individual stimuli. Participants completed 12 runs of the event-related experiment, with each run being composed of 156 TRs. Within each run, 48 images were presented such that after 3 consecutive runs participants had

viewed the entire set of 144 images. Thus, participants viewed 4 complete repeats of the 144 images in total.

Scanning parameters. Participants were scanned on a research-dedicated Siemens 7 T Magnetom scanner in the Clinical Research Center on the National Institutes of Health campus in Bethesda, Maryland. Partial T2*-weighted functional image volumes of the frontal, temporal, and occipital cortices were acquired using a 32-channel head coil (47 slices; 1.6 × 1.6 × 1.6 mm isotropic voxels; 10% interslice gap; TR 2 s; TE 27 ms; flip angle 70°, matrix size 126 × 126; FOV 200 mm). In all scans, oblique slices were oriented approximately parallel to the ventral portion of the prefrontal cortex. In addition, standard MPRAGE (magnetization-prepared rapid-acquisition gradient echo) and corresponding GE-PD (gradient echo–proton density) images were acquired, and the MPRAGE images were then normalized by the GE-PD images for use as a high-resolution anatomical image for the following fMRI data analysis (Van de Moortele et al., 2009).

Functional localizers. During each scan session, an independent functional localizer scan was also collected in each participant to identify scene and face selective regions in ventral temporal and lateral occipitotemporal cortex. The localizer used an on-off design, alternating between 16 s blocks of scene images and blocks of face images presented at 5 × 5° of visual angle. Localizer runs comprised 144 TRs. Participants performed a one-back task, responding to immediate repeats of the same image using a button press.

fMRI data preprocessing. All imaging data were processed using the Analysis of Functional NeuroImages (AFNI) software package (http://afni.nimh.nih.gov/afni, RRID:SCR_005927). Prior to statistical analyses, the functional scans were slice-time corrected and all images were motion corrected to the first image of the first functional run, after removing the appropriate number of 'dummy' volumes (6) to allow for stabilization of the magnetic field. Following motion-correction, data were smoothed with a 2 mm full-width at half-maximum Gaussian kernel.

Functionally defined ROIs. Scene and face selective regions of interest (ROIs) were created for each participant based on the localizer runs. A response model was built by convolving a standard HRF function with the block structure for each run and was correlated to the activation time course. ROIs were generated by thresholding the statistical parametric maps at a threshold of $p < 0.0001$ (uncorrected). Contiguous clusters of voxels (>20) exceeding the defined threshold were defined as scene or face selective. The anatomical locations of these clusters were then inspected to ensure that the current ROIs were consistent with those described in previously published work (Kanwisher, 2010). Our functionally defined face-selective regions included the Fusiform Face Area (FFA) and Occipital Face Area (OFA), and our functionally defined scene-selective regions included the Parahippocampal Place Area (PPA) and the Occipital Place Area (OPA). Ventral early visual areas (vEVC) and dorsal early visual (dEVC) areas (V1–V3) were defined using previously acquired retinotopic field maps from independent participants (Silson et al., 2015, 2016a).

Anatomically defined ROIs. Anatomically defined ROIs were constructed using the Freesurfer image analysis suite, which is documented and freely available for download online (http://surfer.nmr.mgh.harvard.edu/). A ventral temporal cortical (vTC) region was defined using the lower edge of the inferior temporal sulcus as the lateral boundary, extending medially to include the collateral sulcus. Posteriorly, the vTC extended to the edge of the EVC ROIs and anteriorly to the tip of the collateral sulcus. This vTC ROI overlapped with both the functionally-defined FFA and PPA and was drawn to be analogous to the human IT ROI used by Kriegeskorte and colleagues (Kriegeskorte et al., 2008). In addition, a lateral occipitotemporal (lOTC) region was defined extending from the junction of the dorsal and ventral EVC ROIs anteriorly to the superior temporal sulcus, superiorly to the intraparietal sulcus and ventrally to the inferior temporal sulcus. This lOTC ROI overlapped with both the functionally-defined OFA and OPA and also included retinotopic regions such as V3A, LO1 and LO2 (Larsson and Heeger, 2006).

fMRI analysis: event-related data. All 12 functional runs were concatenated and compared to the activation time course for each stimulus condition using Generalized Least Squares (GLSQ) regression in AFNI. In the current paradigm, each image was treated as an independent condition, resulting in 144 separate regressors for each individual stimulus condition, as well as motion parameters and four polynomials to account for slow drifts in the signal. To derive the response magnitude per stimulus, t-tests were performed between the stimulus-specific beta estimates and baseline for each voxel. All subsequent analyses of these data were conducted in Matlab (Mathworks, Natick, RRID:SCR_001622). To derive representational dissimilarity matrices (RDMs), pairwise Pearson's correlations were computed between conditions using the t-values across all voxels within a given ROI (Kravitz et al., 2010, 2011). The resulting RDM for a given ROI was a 144 × 144 matrix representing the pairwise correlations between patterns of activity elicited by each stimulus condition. RDMs were created for each participant, ranked using a tied ranking procedure in which any tied values were assigned their average rank, and then averaged together across participants for each ROI.

Behavior-fMRI comparisons. We calculated full correlations between behavioral judgment RDMs and each of the fMRI derived RDMs (Spearman's *rho*). For all analyses, the behavioral RDMs were based on averages across the maximum number of participants available for that analysis (e.g., all 20 subjects that performed the behavioral experiment for the group-average behavioral judgments; all 10 subjects that performed the behavioral task on Image Set 1 for the group average behavioral RDM of Image Set 1). Statistical significance of between-RDM correlations was determined using fixed-effects stimulus-label randomization tests (Nili et al., 2014). For these tests, a null distribution of between-RDM correlations was obtained by permuting stimulus condition labels of one of the subject-averaged RDMs (e.g., behavioral RDM) 10,000 times, after which the p-value of the observed correlation was determined as its two-tailed probability level relative to the null distribution. In addition, 95% confidence intervals and standard deviations were determined using bootstrap resampling, whereby a distribution of correlation values was obtained by sampling stimulus conditions with replacement (n = 10,000 bootstraps). To correct for multiple testing of the behavioral RDM against the multiple fMRI ROIs, the resulting *p*-values were corrected for multiple comparisons across all ROIs using FDR-correction at $\alpha = 0.05$.

Hierarchical clustering. To reveal higher-order relations between the image categories, the behavioral and fMRI measurements were subjected to hierarchical clustering. To estimate the number of clusters that best described the data, we performed k-means clustering ('kmeans' function implemented in Matlab, 28 iterations) and evaluated the trade-off between number of clusters and explained variance using the elbow method. Using this method, we determined that six clusters optimally described the behavioral data (80% variance explained in each image set). We subsequently performed hierarchical clustering on both the behavioral judgement RDMs and fMRI derived RDMs ('cluster' function in Matlab, method: 'linkage', number of clusters: 6).

Searchlight analysis. To test the relationship between behavioral similarity judgments and activity recorded outside specified ROIs, we conducted whole-brain searchlight analysis. The searchlight analysis stepped through every voxel in the brain and extracted the t-values from a sphere of 3 voxel radius around that voxel (total number of voxels per searchlight sphere = 123), which were then used to compute pairwise correlation distances (1-Pearson's *r*) between each stimulus condition. Analogous to the ROI analyses, the resulting RDMs were correlated (Spearman's *rho*) with the average behavioral RDM. These correlation coefficients were assigned to the center voxel of each searchlight, resulting in a separate whole-volume correlation map for each participant computed in their native volume space. To allow comparison at the group level, individual participant maps were first aligned to their own high-resolution anatomical T1 and then to surface reconstructions of the gray and white matter boundaries created from these T1s using the Freesurfer (http://surfer.nmr.mgh.harvard.edu/, RRID:SCR_001847) 5.3 autorecon script using SUMA (Surface Mapping with AFNI) software

(https://afni.nimh.nih.gov/Suma). Group-level significance was determined by submitting these surface maps to node-wise *t*-tests in conjunction with Threshold Free Cluster Enhancement (Smith and Nichols, 2009) to correct for multiple comparisons, using the CoSMoMVPA toolbox (Oosterhof et al., 2016).

DNN comparisons. Deep convolutional neural networks (DNNs) are state-of-the-art computer vision models capable of labeling objects in natural images with human-level accuracy (Krizhevsky et al., 2012; Kriegeskorte, 2015), and are therefore considered potentially relevant models of how object recognition may be implemented in the human brain (Kriegeskorte, 2015; Yamins and DiCarlo, 2016; Scholte, 2018; Tripp, 2017). DNNs consist of multiple layers that perform transformations from pixels in the input image to a class label through a non-linear mapping of local convolutional filters responses (layers 1–5) onto a set of fully-connected layers of classification nodes (layers 6–8) culminating in a vector of output 'activations' for labels assigned in the DNN training phase. Inspection of the learned feature selectivity (Zhou et al., 2014; Güçlü and van Gerven, 2015; Bau et al., 2017; Wen et al., 2017) show that earlier layers contain local filters that resemble V1-like receptive fields while higher layers develop selectivity for entire objects or object parts, perhaps resembling category-selective regions in visual cortex. The feature representations learned by these DNNs have indeed been shown to exhibit some correspondence with both behavior and brain activity measurements in humans and non-human primates during object recognition (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Güçlü and van Gerven, 2015; Cichy et al., 2016) and scene recognition (Greene et al., 2016; Bonner and Epstein, 2018; Martin Cichy et al., 2017; Groen et al., 2018).

We used the MatConvNet toolbox (Vedaldi and Lenc, 2015) to implement a pre-trained version of an 8-layer deep convolutional neural network (VGG-S CNN) (Chatfield et al., 2014) that was trained to perform the 1000-class ImageNet ILSVRC 2012 object classification task. DNN representations for each individual image in both stimulus sets were extracted from layers 1–5 (convolutional layers) and 6–8 (fully-connected layers) of the network. For each layer, we calculated the Pearson correlation coefficient between each pairwise combination of stimuli yielding one 144×144 RDM per DNN layer. Analogous to the behavior-fMRI analyses, we then calculated Spearman's rank correlations between RDMs derived from DNN layers and RDMs derived from the fMRI and behavioral measurements. Statistical significance was again determined using stimulus-randomization ($n = 10{,}000$ permutations, two-tailed tests). Differences in correlation between individual layers were determined using bootstrap tests ($n = 10{,}000$) whereby the p-value of a difference in correlation between two layers was estimated as the proportion of bootstrap samples further in the tails (two-sided) than 0 (Nili et al., 2014). To correct for multiple testing of several model representations against the same RDM, the resulting p-values were corrected for multiple comparisons across all tests conducted for a given behavioral or fMRI RDM using FDR-correction at $\alpha = 0.05$.

Data availability. All stimuli, models, and the processed behavioral and fMRI data reported in this manuscript are freely available on the Open Science Framework (doi: 10.17605/OSF.IO/MA2RJ, https://osf.io/ma2rj/).

3. Results

The primary aim of this study was to elucidate the representational space of complex naturalistic categories as reflected in human behavior and in neural responses measured with fMRI. We first present analyses examining and comparing the representational structure of each image set estimated from both behavioral similarity judgments and from fMRI responses in visual cortex. We then examine to what extent features derived from a deep neural network (DNN) model can explain the behavioral and fMRI data.

3.1. Comparison of behavioral judgments and fMRI: representational dissimilarity matrices (RDMs)

We first created RDMs based on both the behavioral judgments and fMRI responses, separately for Image Set 1 and Image Set 2. For behavioral judgments, dissimilarities were based on the pixel distances between images in the multi-arrangement similarity task. For fMRI, we focused on the pairwise comparisons of multi-voxel patterns for each stimulus in ventral temporal cortex using a vTC ROI following Kriegeskorte and colleagues ((Kriegeskorte et al., 2008); see Methods). The resulting RDMs are organized alphabetically by category (Fig. 2).

For behavioral judgments, these RDMs exhibit a clear clustering of exemplars within each category for both image sets (Fig. 2A). Participants judged exemplars of the same category as more similar to other exemplars within the same category than to exemplars in different categories (e.g. body parts are more similar to body parts than to mountains). In contrast, there was much less clustering of exemplars for the vTC RDMs, even within category (Fig. 2B). Despite these differences in the RDMs for behavioral similarity judgments and vTC, there were significant, correlations between the two measures (Image Set 1, $rho = 0.06$, 95% CI = [0.02, 0.14], $p = 0.012$; Image Set 2, $rho = 0.07$, CI = [0.03, 0.15], $p = 0.004$), suggesting some similarity in the representation of the images at the image level.

To quantify the extent of category coherence in each image set, we calculated a Category Index as the difference between the average within-category distance and the average between-category distance (Fig. 3A; for a breakdown of the data by category, see Supplementary Fig. 3). For both behavioral judgments and vTC, this Category Index was greater than zero for both image sets (behavior Image Set 1: one-sample t-tests: $t(47) = 41.6$, CI = [0.40, 0.45], $p < 0.0001$; behavior Image Set 2: $t(47) = 44.3$, CI = [0.42, 0.46], $p < 0.0001$; vTC Image Set 1: $t(47) = 5.2$, CI = [0.05, 0.11], $p < 0.0001$; vTC Image Set 2: $t(47) = 6.3$, CI = [0.05, 0.09], $p < 0.0001$), indicating the presence of significant categorical structure in both domains. However, categorization was much stronger for the behavioral judgments compared to vTC (independent samples *t*-test: $t(94) = 29.7$, CI = [0.34, 0.39], $p < 0.001$).

Given the presence of significant categorical structure in both domains, and to directly compare Image Set 1 and Image Set 2, which contained different exemplars for each category, we averaged across exemplars (excluding the diagonal), reducing our 144×144 exemplar-level RDMs to 48×48 category-level RDMs. For both behavioral judgments and vTC there was a strong positive correlation between Image Set 1 and Image Set 2 (behavioral judgments, $rho = 0.64$, CI = [0.55, 0.76], $p < 0.0001$; vTC, $rho = 0.48$, CI = [0.32, 0.67], $p < 0.0001$), indicating that the representational structure in both domains is reproducible across image sets.

Given this reproducibility of representational structure across image sets in both behavior and vTC, we averaged across sets to compare the representational space at a category-level between behavior and vTC (Fig. 3B and C). As with the exemplar level, there was a significant correlation between behavioral judgments and vTC ($rho = 0.10$, CI = [0.02, 0.32], $p = 0.019$), suggesting some similarity in the representational structure. However, this correlation was weaker than the relationship between Image Set 1 and Image Set 2 within behavior and vTC separately (Fisher' r to z transformation: behavior-vTC correlation vs. behavior-behavior Image Set correlation: $z(48) = 3.1$, $p = 0.002$ (two-tailed); behavior-vTC correlation vs. vTC-vTC Image Set correlation: $z(48) = 2.0$, $p = 0.045$ (two-tailed)). The within-domain correlations provide an estimate of the maximum correlation that could be achieved across domains. Therefore, the much greater correlation within-domain than across domains suggests that each domain is only explaining a small proportion of the explainable variance in the other. A similar pattern of within- and between domain correlations was observed when looking at the individual participant level (see Supplementary Fig. 4).

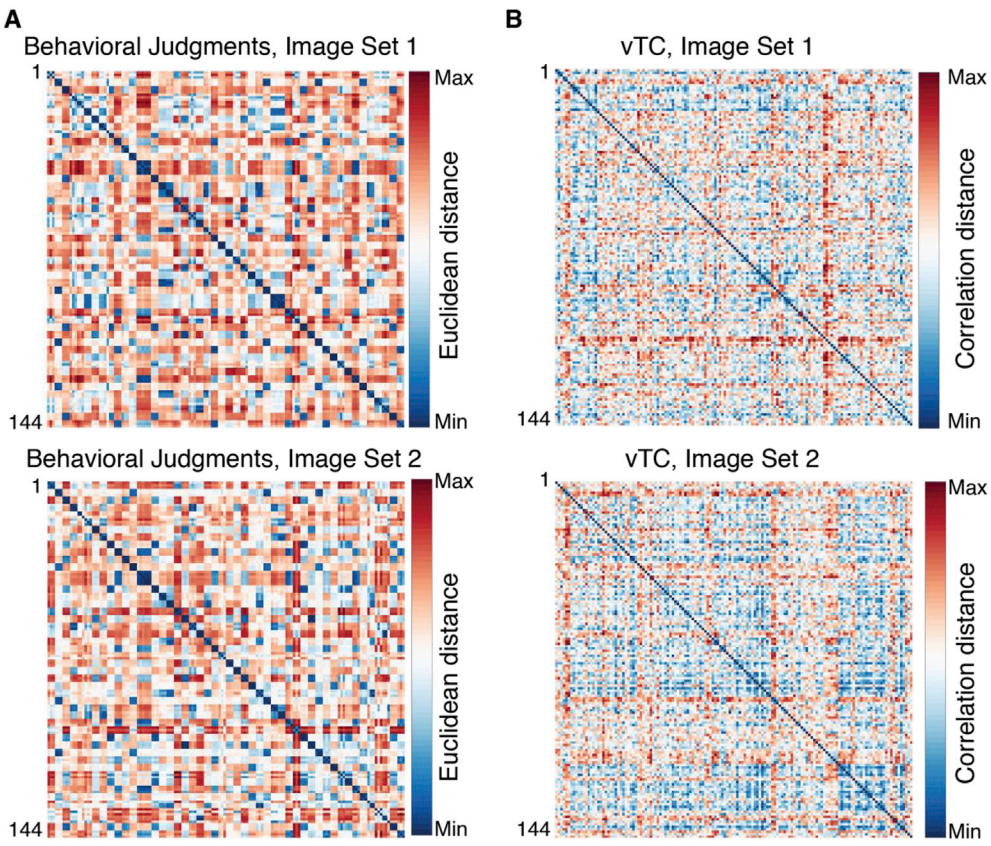

Fig. 2. Representational dissimilarity matrices for Image Set 1 and Image Set 2. Matrices show comparisons for all 144 images grouped alphabetically by category (3 images per category, same order as Fig. 1). A) Behavioral dissimilarity was measured as the Euclidean distance between pairs of images in the multi-arrangement task. Clustering-by-category is evidenced by the appearance of 3×3 exemplar 'blocks' exhibiting low dissimilarity along the diagonal. B) fMRI dissimilarity was measured as 1 minus the pairwise correlation between the pattern of response to images in vTC. There is some clustering-by-category present, but it is less evident than for the behavioral judgments.

3.2. Structure of category representations: hierarchical clustering

Despite the significant correlations between the behavioral and vTC RDMs at the group level, the structure of the RDMs appear to be quite different, and the smaller correlation between than across experimental measures suggests there may be differences in the representational spaces. To investigate the nature of the category representational structure, we conducted hierarchical clustering analyses (see Materials and Methods). For behavioral similarity judgements, a group of clear and intuitively meaningful clusters emerged, including clusters that appear to reflect 'urban landscapes', 'transportation', 'humans', 'household items', 'animals/insects', and 'natural scenes' (Fig. 4A, left). The first branching point in the dendrogram separates animals/insects and natural scenes from all other categories. Thus, animal categories (e.g. farm animals, wild animals) were not grouped with people (i.e. by animacy), but rather were grouped closest to natural objects and scenes (e.g. fire, flowers, beaches). Human categories (e.g. adults, older adults, kids, sports, and body parts) were grouped most closely to people-related objects (e.g. human food, airplanes, trains, bikes) and people-related places (e.g. living rooms, kitchens). These results suggest that behaviorally, participants tended to group images into manmade (including humans) and natural categories (including animals).

In contrast, however, hierarchical clustering based on data derived from vTC revealed a relationship between categories that is much harder to characterize (Fig. 4B, left). In general, it appears that some categories containing stimuli with faces and/or bodies (e.g. wild animals, pets, dolls, older adults, kids, adults) were represented as similar to one another and distinct from all other categories in vTC, a division that is reflected in the first branching point of the dendrogram. However, there is not a clean grouping of images containing faces and/or bodies from all others since some categories containing faces or bodies (e.g. farm animals, masks) were not contained in the same cluster. In terms of a possible animate/inanimate distinction, it is clear that many animate categories (e.g. lizards/snakes, spiders) were clustered with inanimate categories (e.g. food, flowers, boats, etc.).

Applying the hierarchical clustering orders to the behavioral and vTC RDMs (Fig. 4A and B right) highlights the differences between the behavioral and vTC RDMs. When the behavioral clustering order is applied to the vTC RDM, very little structure is present except for the grouping of the categories of kids, adults and older adults, which were relatively more similar to each other than any other categories except for farm animals, wild animals and pets. This suggests some similarities in the representation of kids, adults and older adults between behavior and vTC. When the vTC clustering order is applied to the behavioral RDM, many of the clusters in the behavioral data become fragmented, but some groupings remain. For example the grouping of older adults, kids and adults is clear as well as that of farm animals, butterflies and birds.

In sum, the hierarchical clustering reveals no evidence for a separation of animate and inanimate categories in either the behavioral or the vTC RDM. Moreover, we observe clear differences in the representational structure of the behavioral and vTC RDMs, with more discrete clustering in the behavioral compared to the fMRI domain. The one clear consistency between the behavioral and vTC RDMs is the grouping of the kids, adults and older adults categories. In the next section, we consider whether the differences between the behavioral and vTC RDMs reflect the particular ROI chosen for the fMRI data.

3.3. Beyond the vTC ROI

To investigate whether the weak relationship observed between the behavioral and vTC RDMs reflects the *a priori* choice of ROI, we identified a number of other ROIs in visual cortex and conducted a series of exploratory analyses to determine if any of these regions are more closely correlated with the representational structure that emerged in the behavioral similarity judgments.

First, we defined a series of new ROIs using either independent

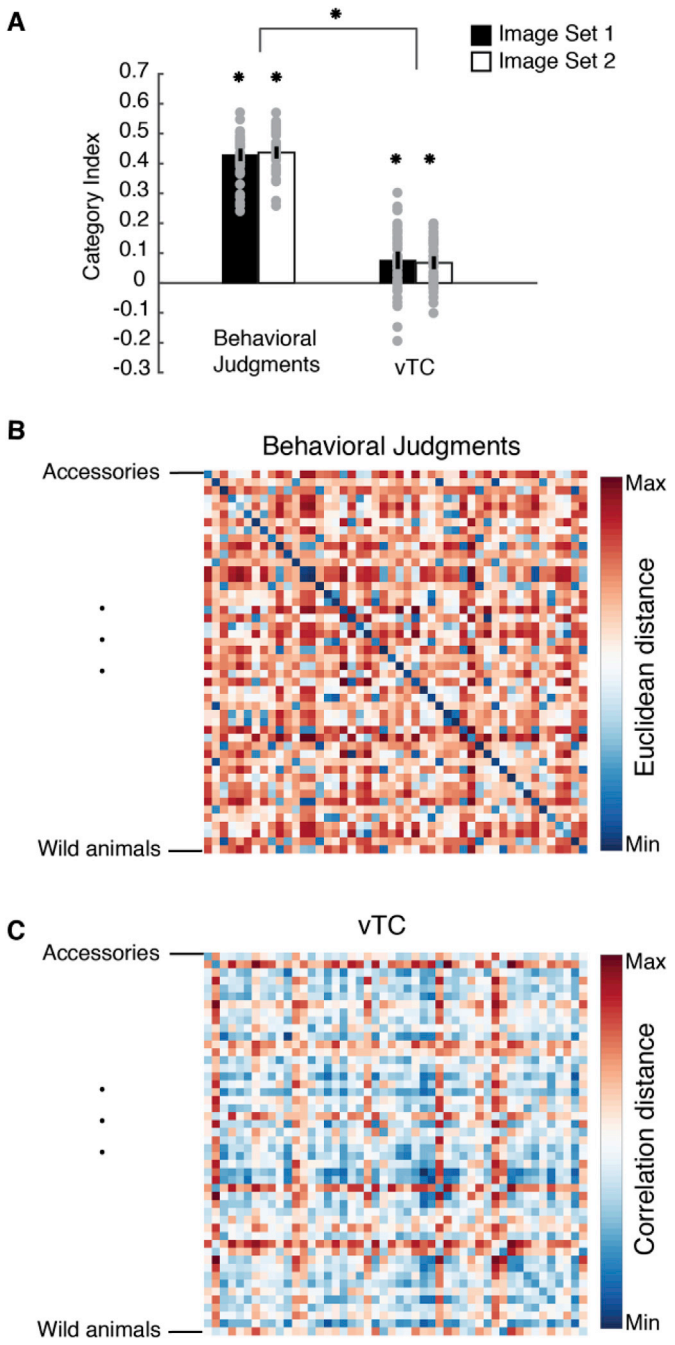

Fig. 3. Category representations. a) Category indices for vTC and behavioral similarity judgements calculated as the difference between the average within-category and between-category distances, averaged across categories. Gray dots indicate indices for each category separately (for a full break down by category, see Supplementary Fig. 3). Error bars indicate 95% confidence intervals estimated from a one-sample *t*-test. * = $p < 0.001$. b), c) RDMs averaged by category for behavioral similarity judgements and fMRI responses in vTC. Categories are ordered alphabetically in the matrices.

functional localizers and anatomical constraints (see Methods and Fig. 5A). In particular, we examined i) a high level visual region in lateral occipitotemporal cortex (lOTC), analogous to the vTC, incorporating face-, scene-, and object-selective regions, ii) functionally-defined category-selective regions, including both face-selective (FFA and OFA) and scene-selective (PPA and OPA) regions in ventral temporal and lateral occipital cortex, respectively and iii) early visual cortex (EVC) ROIs (combining V1–V3) subdivided into a dorsal (dEVC) and ventral (vEVC) division. We compared the RDMs for each ROI across Image Set 1 and Image Set 2 and also correlated them with the RDM for behavioral judgments.

The diagonal of the ROI comparison matrix (Fig. 5B) indicates the reliability of the representational structure across image sets and participants. There are clear differences in the strength of the correlations for the different ROIs. In general, reliability was higher for the ventral compared to the dorsal ROIs (vTC vs. lOTC, vEVC vs. dEVC, FFA vs. OFA, PPA vs. OPA). Further, the representational structure differed across ROIs. For example, the representational structure in the EVC ROIs was very different from that observed in the higher-level ROIs. The vTC ROI, which we used in our analyses so far, varied in its relationship with the other ROIs, showing highest similarity with PPA and lOTC, and lowest with dEVC and vEVC.

For behavior, we compared the RDM for each ROI with the behavioral similarity RDM. PPA showed the strongest correlation ($rho = 0.22$, CI = [0.08, 0.46], $p < 0.0001$) followed by OPA ($rho = 0.16$, CI = [0.07, 0.38], $p < 0.0001$) (Fig. 5C), although these correlations were again much weaker than the correlation of the PPA RDM across image sets ($rho = 0.41$, CI = [0.28, 0.59], $p = 0.0002$). The weakest relationship was observed for FFA, which actually showed a trend towards a negative correlation ($rho = -0.09$, CI = [-0.13, 0.10], $p = 0.06$), despite showing a strong positive correlation across image sets ($rho = 0.49$, CI = [0.38, 0.63], $p < 0.0001$).

Second, we conducted an exploratory searchlight analysis to examine any other brain areas that might show a relationship to the representational structure of the stimuli that emerged in behavioral similarity judgments. Our slice prescription included all of occipital, temporal and parietal cortex, but not frontal regions. The strongest brain-behavior correlation emerged in areas corresponding to scene-selective regions PPA and OPA (Fig. 6), as well as a medial parietal region that seems to correspond to a third scene-selective region (medial place area, MPA, also referred to as retrosplenial complex, RSC (Epstein, 2008; Silson et al., 2016b)).

Taken together, these data indicate the strongest relationship between the representational structure of behavioral similarity judgments and fMRI responses is in scene-selective cortex, particularly PPA, followed by OPA, while the weakest relationship was observed for FFA. This could be considered surprising, given that the one clear consistency between the behavioral judgments and fMRI responses in vTC (a large ROI that encompasses both PPA and FFA) appeared to reflect a grouping of the adults, kids and older adults categories, which are image categories that FFA responds strongly to, but PPA does not. To further explore the origin of this correspondence, we next examined the representational structure in PPA and FFA and their relation with the behavioral dissimilarity in more detail.

3.4. Representation of human categories in PPA and FFA

Hierarchical clustering (Fig. 7) indicated that both PPA and FFA contained an early branching of a cluster that included adults, kids and older adults, similar to the larger vTC ROI. However, in PPA, this cluster also included body parts, while in FFA this cluster also included sports (which typically contained people) and dolls. Further, inspection of their respective RDMs (Fig. 8A) revealed some clear differences in representational structure. While for both FFA and PPA the categories of adults, kids and older adults showed strong dissimilarity with most other categories (resulting in them being grouped in a separate cluster in both ROIs), in FFA these categories were also similar to one another, as well as to pets, wild animals and farm animals. In contrast, PPA showed no such within-cluster similarity of these categories. Instead, PPA exhibited high similarity for urban scenes such as houses, cityscapes and churches, categories that in turn were highly dissimilar from one another in FFA.

This difference between the PPA and FFA RDMs was further highlighted when the correlation between PPA or FFA and behavioral judgments was computed for each category separately (Fig. 8B). High

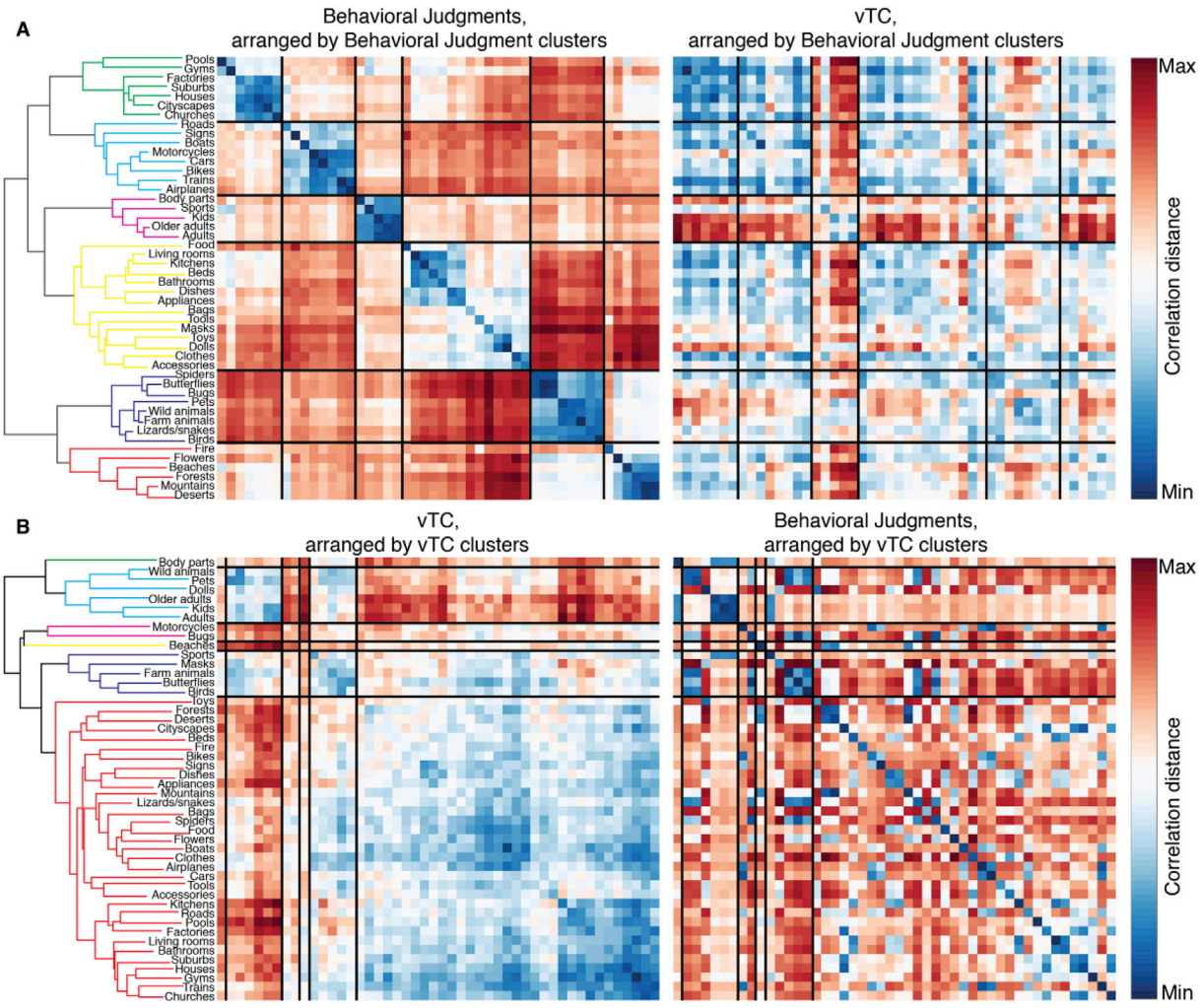

Fig. 4. Hierarchical clustering of behavioral and vTC RDMs. A) Hierarchical clustering of behavioral similarity judgments. RDMs for behavior (left) and vTC (right) arranged in the behavioral dendrogram order. B) Hierarchical clustering of vTC dissimilarity. RDMs for vTC (left) and behavioral judgments (right) arranged in the vTC dendrogram order. Dendrograms are colored according to the top six clusters and the black lines on the RDMs show the boundaries between these clusters.

correlations indicate that the category was similarly represented in the fMRI and behavioral RDM, while low or negative correlations indicate differences in the representational structure. For PPA, most categories showed a positive correlation, with the strongest correlations for urban landscapes such as factories, houses and cities. The lowest correlations were observed for categories containing humans or faces such as adults, kids, masks and dolls. In contrast, in FFA, most of the correlations were negative, indicating a striking difference in the representational space for most categories. The strongest positive correlations were observed for categories containing people and for animals. Collectively these analyses suggest that PPA and FFA each capture different aspects of the behavioral similarity judgements.

In sum, comparisons of regions beyond the vTC ROI suggest that representational structure was most reliable for ventral regions, with clear differences in representational structure between regions. Out of all ROIs examined, scene-selective regions correlated best with behavior, and this observation was supported by the searchlight results. However, relative to the reproducibility within the fMRI domain, the magnitude of the fMRI-behavior correlations remained relatively weak. The separation of the kids, adults and older adults categories that we observed for vTC was evident in hierarchical clusters obtained for both PPA and FFA. However, for PPA, the correlation with behavior was driven by non-face categories, while FFA only correlated weakly with behavior for those categories and exhibited limited correspondence for other categories. Collectively, these results suggest that neither ROI fully captured the representational structure reflected in the behavioral judgments. To better understand what is being represented in behavioral judgements and fMRI responses, we next considered a third domain of representation: computational modeling.

3.5. *DNN comparisons with fMRI responses and behavioral judgments*

In light of previous reports showing a correspondence between DNNs and both behavioral judgments and brain activity measurements in humans and non-human primates, we next examined to what extent DNN representations were able to explain the representational structure observed in our current data. Given the discrepancy between our fMRI and behavioral measurements, we were particularly interested to determine which of the two domains corresponded more strongly with the DNN representations.

We created RDMs based on DNN representations for individual layers of an 8-layer, off-the-shelf pre-trained DNN (see Materials and Methods), separately for Image Set 1 and Image Set 2. Dissimilarities were calculated as the correlation distances between the vectorized responses across all units within a given layer. Similar to the behavioral and fMRI measurements described above, representational structure within each DNN layer (Fig. 9A) was reproducible across image sets, increasing gradually from lower to higher layers (Image Set 1 versus Image Set 2, all

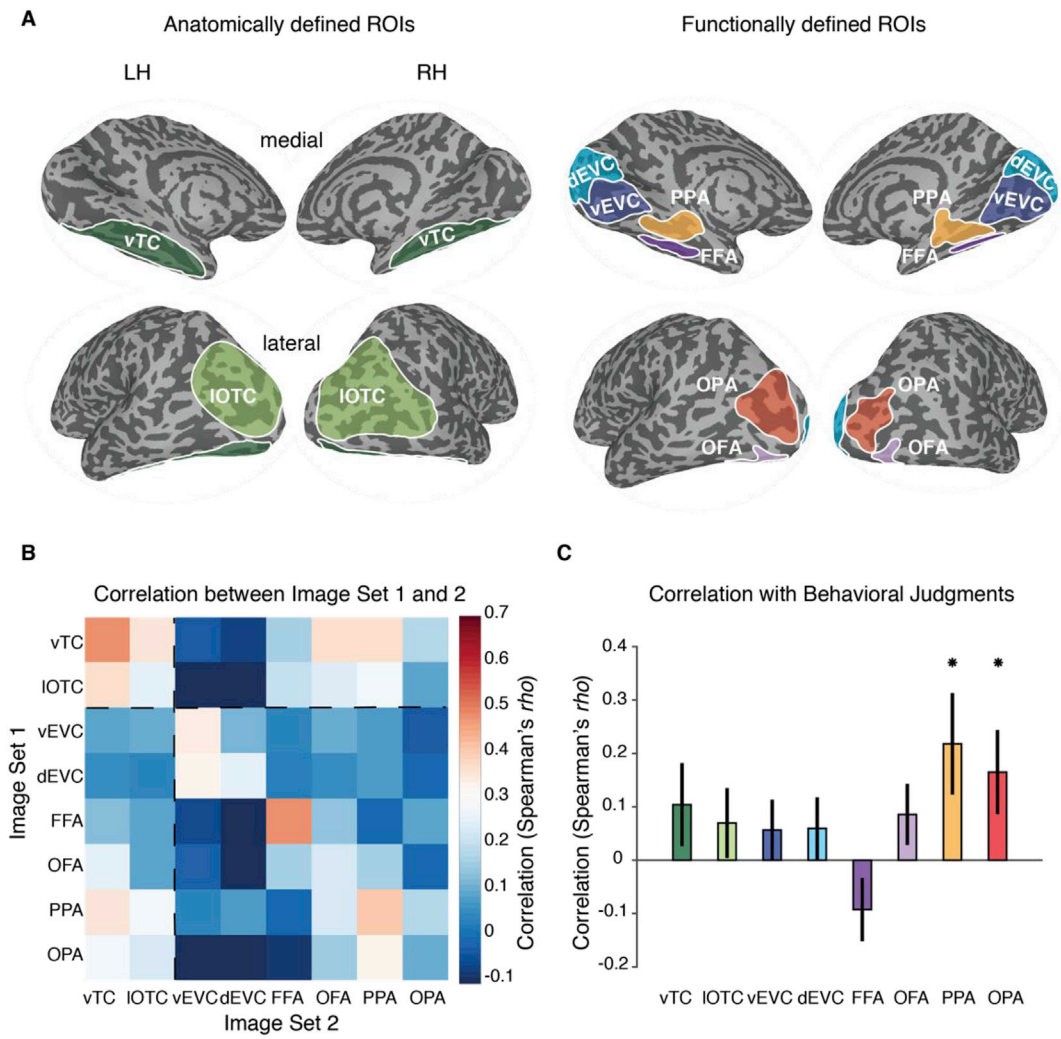

Fig. 5. Comparison of multiple visual cortical ROIs. A) Anatomically (left) and functionally defined (right) ROIs. Anatomical and category-selective ROIs were defined in each individual participant. Early visual cortex ROIs were defined at a group-level in an independent set of participants. B) Correlation between the RDMs for each region of interest. Correlations are computed between participants viewing Image Set 1 and those viewing Image Set 2. ROIs included high-level visual cortex on the ventral (vTC) and lateral (lateral occipitotemporal cortex, lOTC) surfaces, dorsal and ventral early visual cortex (dEVC, vEVC), face-selective (OFA, FFA) and scene-selective (OPA, PPA) cortex. Correlations within a ROI were higher on the ventral compared to the lateral/dorsal cortex for all pairs of regions. C) Correlation between the average behavioral RDM and the RDM for each ROI. * Significant correlations relative to zero (two-tailed, FDR-corrected) as assessed with a permutation test (n = 10,000). Error bars reflect the standard deviation of the bootstrap distribution of correlation values. The strongest correlation was observed in PPA and the weakest in FFA. Note that the multiple comparisons correction renders the correlation between behavior and vTC reported in our earlier analyses insignificant.

rho = [0.21, 0.62], all $p < 0.0001$). For comparisons with representational structure in the behavioral judgments and fMRI, responses we averaged the RDMs across the two image sets separately for each layer. We then compared the representational structure of each layer with the RDMs for behavioral judgments and a number of fMRI ROIs (Fig. 9B).

For behavior, we observed a consistent correlation with the DNN that gradually increased with higher layers, culminating in the highest correlation for layer 8 (*rho* = 0.56, CI = [0.46, 0.69], $p < 0.0001$). In contrast, the highest correlation with PPA was found for layer 5 (*rho* = 0.55, CI = [0.44, 0.68], $p < 0.0001$); while its correlation also gradually increased from layer 1 to 5, higher layers did not differ significantly from layer 5. A similar pattern of results was observed for the larger vTC ROI (highest correlation with layer 5: *rho* = 0.44, CI = [0.32, 0.60], $p < 0.0001$). In contrast, none of the DNN layers exhibited a significant correlation with FFA, whose correlations instead appeared to trend negatively (all *rho* = [-0.18, −0.01], all $p > 0.05$), similar to the relationship between FFA and behavior.

These results demonstrate that higher-level DNN representations are reproducible across image sets and, surprisingly, are correlated with *both* the behavioral and the brain measurements in PPA and vTC, with relatively high maximal correlations for both domains (around *rho* = 0.55). However, behavioral and fMRI representational similarity differed in terms of which layer correlated more strongly. For behavioral judgments, higher layers invariably resulted in increasing correspondences with behavior, all the way to the top-most layer (layer 8) that is closest to the output (see Supplementary Fig. 5 for a more detailed comparison of the representational structure of layer 8 with behavior). In contrast, correlations with fMRI measurements in high-level cortex regions increased up to mid-level layer 5, only to plateau or even decrease again for subsequent layers.

This result suggests that additional computations carried out in the fully-connected layers (6–8) are important to explain human behavioral judgments, but not fMRI responses, which map more strongly onto representations contained in the mid-to-high-level convolutional layers.

4. Discussion

We compared the representational similarity of behavioral judgments

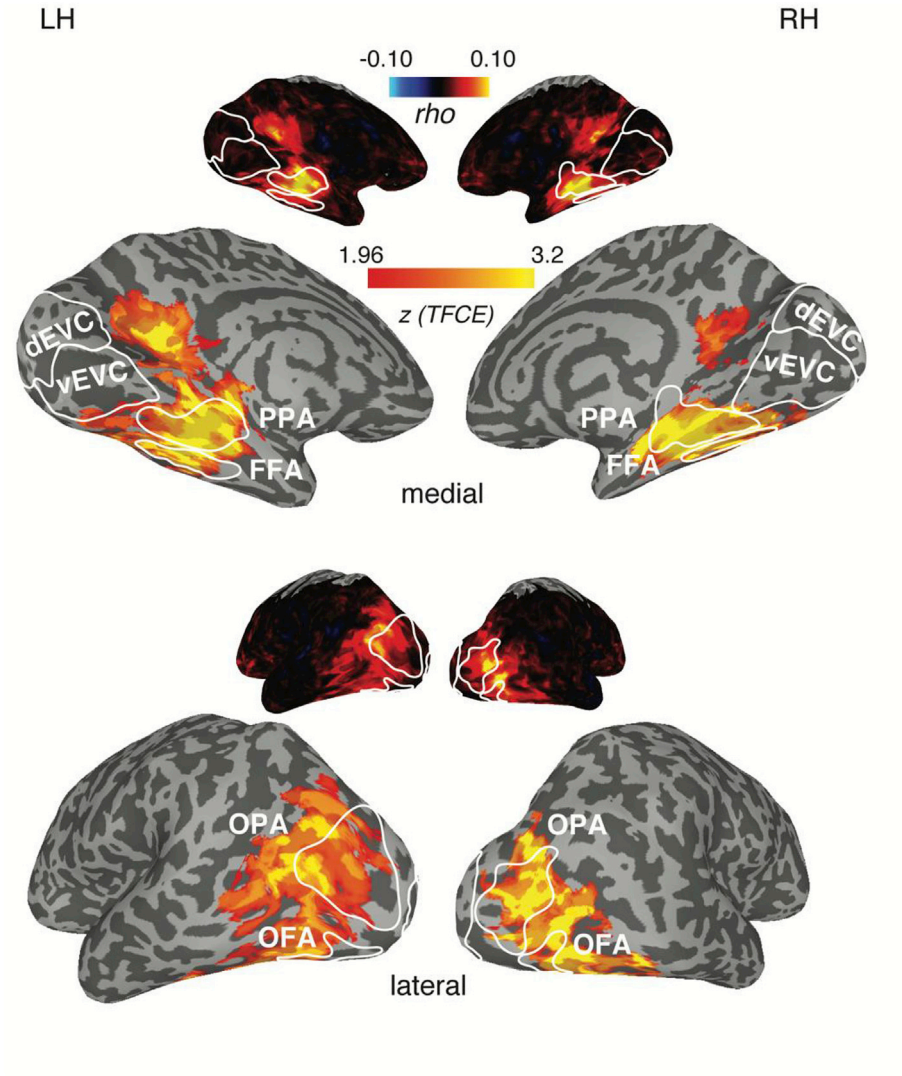

Fig. 6. Behavioral RDM searchlight results. The strongest correlations with the behavioral RDM were observed in scene-selective regions OPA and PPA. There was also a strong correlation in medial parietal cortex that likely corresponds to a third scene-selective region, MPA (medial place area). Small brains show the unthresholded correlation values and large brains are cluster-corrected for multiple comparisons using Threshold-Free Cluster Enhancement (thresholded on $z = 1.94$, corresponding to two-sided $p < 0.05$). Group-level results are overlaid on the freesurfer reconstruction of one example participant, with the corresponding functionally-defined ROIs highlighted in solid white lines.

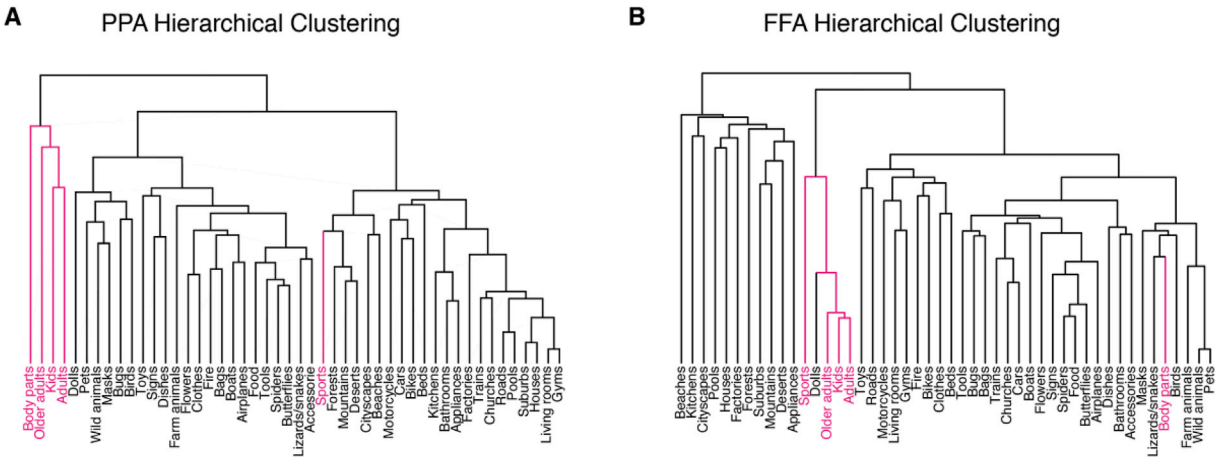

Fig. 7. PPA versus FFA: hierarchical clustering. A) Hierarchical clustering of representational dissimilarity in scene-selective PPA indicated the presence of a face- and body-selective cluster (first branch) containing the categories adults, kids and older adults, as well as body parts. B) Hierarchical clustering of face-selective FFA indicated a face-selective cluster (second branch) containing adults, kids and older adults, as well as sports (which typically included people) and dolls. Branches highlighted in magenta indicate cetegories contained within the 'human' cluster derived from behavioral judgments (Fig. 4A).

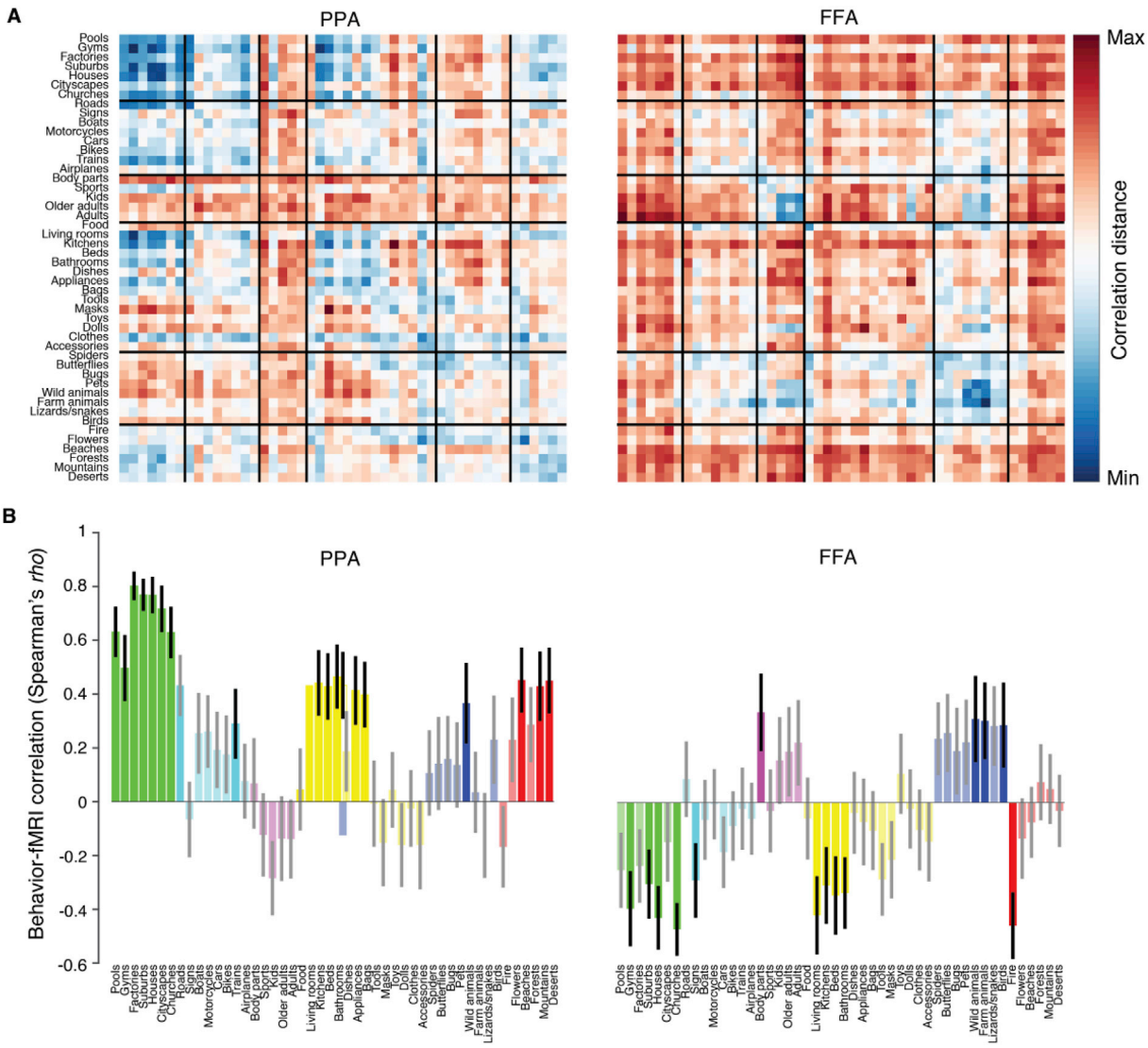

Fig. 8. PPA versus FFA: RDMs and individual category correlation with behavior. A) RDMs of PPA and FFA arranged in the behavioral clustering order. Superimposed black lines indicate the clusters derived from the behavioral judgments RDM (see Fig. 4A). B) For each category, correlations were computed between PPA (left) or FFA (right) dissimilarity and behavioral dissimilarity (Spearman's *rho*). Individual correlations are color-coded by the clusters derived from behavioral judgments. Significant correlations are depicted as opaque bars, while non-significant correlations are transparent. Significance was assessed using a permutation test with 10,000 permutations per category ($p < 0.05$, two-tailed). Error bars reflected the standard deviation of the bootstrap distribution of correlations (10,000 bootstraps).

with those derived from fMRI measurements of visual cortex for a set of naturalistic images drawn from a range of object and scene categories. Despite a significant correlation between the fMRI and behavioral measurements, this correlation was much weaker than within each experimental measure separately, and there were clear differences in the overall geometry of the representational spaces. While the behavioral data revealed a broad distinction between manmade (including humans) and natural (including animals) content, with clear sub-groupings of categories sharing conceptual properties (e.g., transportation: roads, signs, airplanes, bikes), the fMRI data largely reflected a division between images containing faces and bodies (e.g. kids, adults, older adults, body parts) and other types of categories, with sub-groupings that were very heterogeneous. This discrepancy was not due to the specific cortical regions chosen, and even the region showing the strongest correlation with behavior (scene-selective PPA) exhibited quite distinct representational structure from that observed for behavioral judgments. An off-the-shelf DNN appeared to explain both the behavioral and fMRI data, yet the behavior and fMRI data showed maximal correspondences with different layers, with fMRI responses mapping more strongly onto middle levels of representation compared to behavior. Collectively, these results demonstrate that there is limited correspondence between multi-voxel responses in visual cortex and behavioral similarity judgments and highlight the importance of probing representational structure beyond just correlation of the dissimilarity matrices. Below, we discuss three potential explanations for this divergence.

4.1. *Visual versus conceptual information*

One possibility is that while the fMRI data reflect the visual properties of the stimuli, behavioral similarity judgments reflect conceptual structure that goes beyond those visual properties. Such a view is consistent with prior studies demonstrating that low-level visual properties contribute to responses in high-level regions of visual cortex (Watson et al., 2017; Groen et al., 2017). Our comparison with the DNN representations seem to support this suggestion, with fMRI most related to layer 5 and behavior corresponding most strongly to layer 8, consistent with prior studies reporting a peak correlation between scene-selective cortex and layer 5 in similar networks (Bonner and Epstein, 2018;

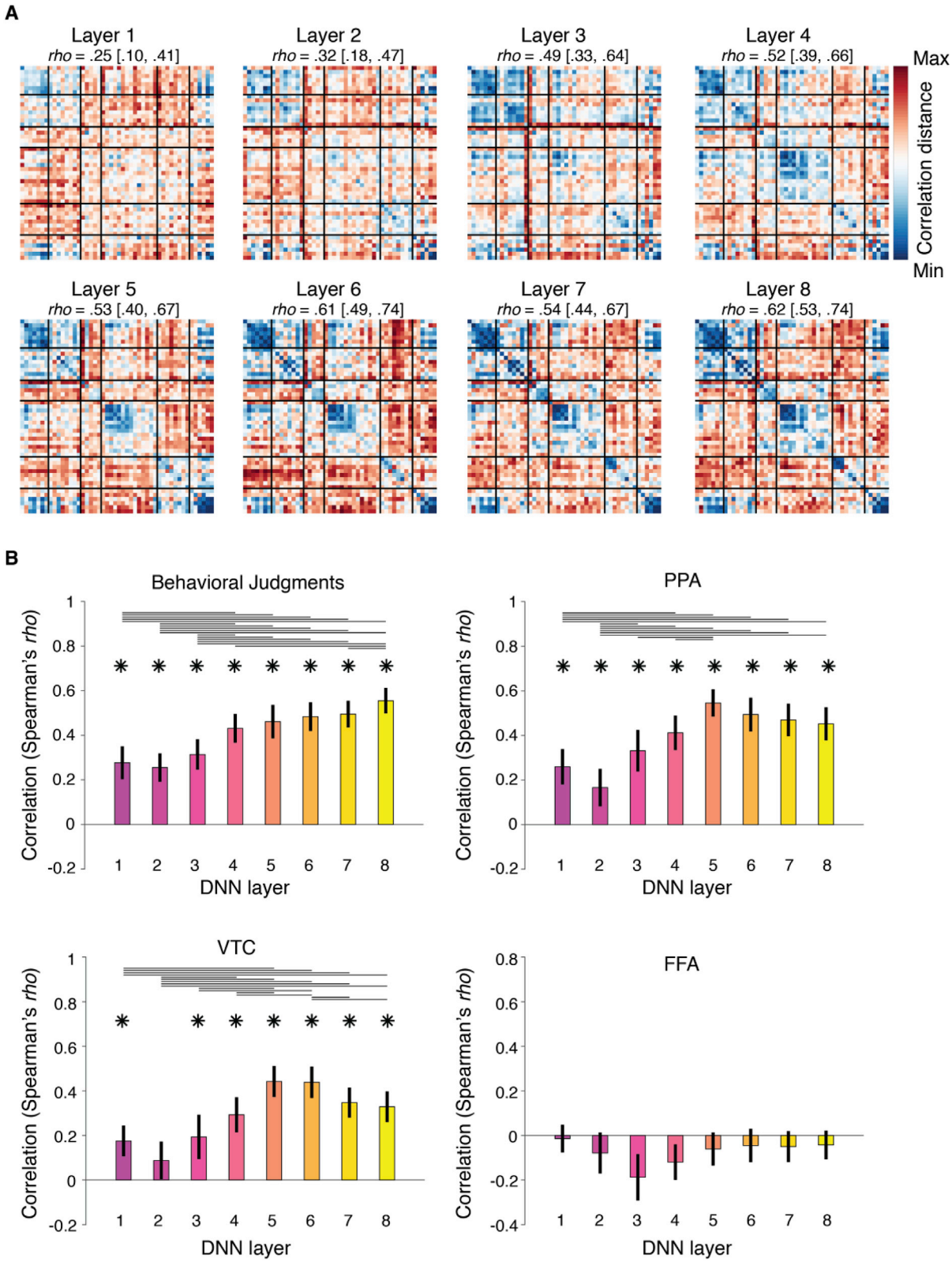

Fig. 9. DNN representations correlate with brain and behavior. A) RDMs (correlation distances) for each of the 8 layers of the DNN, ordered based on the hierarchical clustering of the behavioral RDM. Superimposed black lines indicate the cluster derived from the behavioral judgments RDM (see Fig. 4A). The between set correlation values above each RDM (*rho* [95% CI]) increase with layer number, reflecting increased reproducibility of representational structure for higher DNN layers. **B)** Correlation of each individual layers with behavior, vTC, PPA and FFA. * significant correlations (FDR-corrected) relative to zero (two-tailed) as assessed with a randomization test (n = 10.000). Horizontal lines indicate significant differences (FDR-corrected) between correlations (two-tailed) as assessed with bootstrapping (n = 10.000). Error bars reflect the standard deviation of the mean correlation, obtained via a bootstrapping procedure (see Methods).

Groen et al., 2018; Khaligh-Razavi and Kriegeskorte, 2014). The type of DNN layer may be an important factor as layers 1–5 are convolutional and contain features that can be visualized (Zeiler and Fergus, 2014) and are still spatially localized in the image. In contrast, layers 6–8 perform a mapping of those features onto the class labels used in training. Thus the later DNN layers contain a potentially more fine-grained representation that better matches behavior of human observers, while the fMRI responses correspond to an earlier stage of processing where visual features relevant for categorization are represented at a coarser level.

Others have suggested, however, that hierarchical visual models (e.g. HMax, DNN) do not capture semantic or conceptual information and that an additional level of representation is required (Clarke and Tyler, 2014; Clarke et al., 2015; Devereux et al., 2018). However, this view tends to discount the covariance between visual features and conceptual properties as well as co-occurrence statistics (e.g. a banana and an orange are much more likely to occur in an image together than a banana and a motorcycle). Indeed, the correspondence we observed between the higher levels of the DNN and behavioral similarity judgments, which appear to reflect fine-grained groupings of conceptually-related stimuli, suggests that a significant amount of conceptual information can be captured by a hierarchical visual model, although there are still differences in the overall representational structure (Supplementary Fig. 5).

While we focused on visual cortex, it has been reported that conceptual representations are reflected beyond visual cortex in perirhinal cortex (Devereux et al., 2018; Martin et al., 2018). However, our searchlight analysis demonstrated the strongest correlations between fMRI and behavioral similarity measures in scene-selective regions and did not highlight perirhinal cortex. Our slices included occipital, temporal and parietal cortices but not prefrontal cortex, so it is possible that a stronger correspondence between the fMRI and behavior could emerge there.

4.2. Organization of representations in the cortex

Another reason for the divergence between the behavioral and brain measurements may be that we failed to adequately capture the representational organization of categorical information in visual cortex. In this study we compared behavioral similarity judgments with representations in regionally-localized brain regions using multi-voxel patterns. In this context, there are two important factors to consider, namely i) the scale and ii) the distribution of information representation in the cortex.

First, multi-voxel patterns may primarily reflect the large-scale topography of cortex rather than more fine-grained representations (Freeman et al., 2011). In high-level visual cortex, there are large-scale differences across the vTC reflecting the categorical distinction between faces and scenes that overlap with an eccentricity gradient (Hasson et al., 2002) and variation according to the real-world size of objects (Konkle and Oliva, 2012). These considerations are consistent with the general grouping we observed in the fMRI data that seemed to reflect a separation of images with faces and bodies from all other images. An alternative approach to using multi-voxel patterns is to model feature-selectivity at the individual voxel level (Naselaris et al., 2011). While this approach might be more sensitive to more fine-grained selectivity, it is striking that studies using this approach have primarily revealed smooth gradients across visual cortex that largely seem to reflect the large-scale category-selective organization (Huth et al., 2012; Wen et al., 2018) with evidence for a limited number of functional sub-domains (Çukur et al., 2013, 2016).

Second, the behavioral similarity judgments revealed apparent conceptual groupings that likely reflect multiple dimensions on which the images could be evaluated. A strong correspondence between a localized cortical region and the behavioral similarity judgments would suggest that all those dimensions are represented in a single region (i.e. a 'semantic hub'; (Patterson et al., 2007)). However, we found no such region in our searchlight analysis, suggesting that if it does exist, it likely lies outside of visual cortex. Alternatively, conceptual knowledge may be distributed across multiple regions with each representing specific object properties (Martin, 2016) and there is some fMRI evidence for distributed semantic representations (Huth et al., 2012). However, we also failed to observe a good correspondence with behavior in our vTC ROI, which include a large proportion of high-level visual cortex. While it is possible that some differential weighting of the response across this region may have led to a better fit with the behavioral response, this possibility only further highlights the difficulty in mapping between the response of high-level visual cortex and behavior.

4.3. Impact of task and stimuli

Finally, it is worth considering the nature of the stimuli, the behavioral task and the differences between the behavioral and fMRI tasks (see also (Bauer and Just, 2019)). First, we presented stimuli on naturalistic backgrounds to more closely reflect our everyday experience. However, this may have encouraged participants to sometimes base their behavioral judgments on the backgrounds, rather than the objects, and this may partly explain why the strongest correlations between behavioral and brain measures are in scene-selective regions. Further, if the backgrounds are processed to different degrees during the behavioral and fMRI sessions, this may account for the difference in representational structure. However, we do not think the use of naturalistic backgrounds can explain the divergence we observed between fMRI and behavior. In our stimuli, objects were always presented on congruent backgrounds and there is typically a strong association between the nature of the background and the nature of the object, as well as the visual properties of each. Further, in a preliminary pilot experiment (data not published) for a prior publication (Bankson et al., 2018), we tested the impact of backgrounds on the behavioral arrangement of a set of 84 objects. We found very similar representational structure between the two sets of stimuli (Spearman's $rho = 0.62$), suggesting a limited impact of the naturalistic backgrounds. Finally, in a recent study (Cichy et al., 2019) significant correlations between behavioral performance in the arrangement task and representations in high-level visual cortex were reported even when participants were told to focus on the backgrounds. Specifically, in addition to 'free' arrangement, Cichy and colleagues also tested arrangement tasks in which participants were told explicitly to use shape, color, function or background. All of the resulting RDMs were significantly correlated with each other and with the free arrangement RDM, and for the shape, background and free arrangement RDMs, the strongest correlations with brain data were co-localized in high-level visual cortex, in areas consistent with those we report. Thus, while it will be important for future work to address the relative contribution of background to our understanding and representations of visual stimuli, it seems unlikely that the presence of the naturalistic backgrounds can easily explain the differences in representational structure between brain and fMRI data that we observed.

Second, in the behavioral task, participants were simply told to arrange the stimuli according to similarity without directing them to use any particular dimension to evaluate the stimuli. We chose this task because prior studies have reported a close correspondence between performance on this task and fMRI measures in vTC (Kriegeskorte et al., 2008; Mur et al., 2013; Cichy et al., 2019). One concern, however, is that because of the lack of constraints, participants could have used any from a wide range of possible dimensions to evaluate the stimuli and the particular dimensions chosen might vary across participants. However, the correlations between individual participants (Supplementary Fig. 4), even when using different stimulus sets, suggests that participants did not use markedly different strategies. An alternative approach would have been to give participants specific instructions for how to consider similarity (e.g. shape similarity, semantic similarity), but then we would simply be recovering the dimensions participants were explicitly told to focus on in the results rather than revealing participants implicit understanding of the stimuli, which we assume reflects the integration of multiple possible dimensions for each stimulus. Further, as noted above,

similar results have been reported for free arrangement compared to explicit arrangement tasks, both in terms of behavior and in terms of the correlation with fMRI data in vTC (Cichy et al., 2019). Thus, while there are a lot of different behaviors that could be reflected in the free arrangement task, to some extent they may depend on the same underlying brain representational space.

Third, the behavioral task required participants to compare simultaneously presented stimuli and make explicit similarity judgments, but an unrelated fixation cross task was performed during fMRI. It is thus possible that during fMRI participants processed the images differently, resulting in a different representational space (Mur et al., 2013) and a more explicit and involved fMRI task might have yielded more similar representations across tasks. However, we used very similar tasks to those used in prior studies that reported a closer correspondence between fMRI of vTC and behavioral similarity judgments (Kriegeskorte et al., 2008; Mur et al., 2013; Cichy et al., 2019). Therefore, we postulate that the reduced correspondence we observed in our study likely reflects the broader sampling of object and scene categories, particularly for inanimate categories. Further, while task has been reported to have a strong impact on behavioral representations (Schyns and Oliva, 1999; Harel and Bentin, 2009; Bracci et al., 2017a), fMRI studies have found limited effects of task on representations in vTC (Harel et al., 2014; Bracci et al., 2017a; Groen et al., 2018; Hebart et al., 2018). Instead, task effects appear to be much more prevalent in parietal and frontal regions (Erez and Duncan, 2015; Bracci et al., 2017a; Vaziri-Pashkam and Xu, 2017). In fact, the relative inflexibility of representations in vTC compared to behavior further highlights the difficulty in directly mapping between them.

4.4. Representation of animacy

One striking aspect of our results is that contrary to previous work (Kriegeskorte et al., 2008; Naselaris et al., 2012; Mur et al., 2013; Sha et al., 2015) we did not observe a clear separation of animate vs. inanimate categories in either behavioral or fMRI representational similarities. Instead, in behavior, images were initially grouped according to a broad division between man-made (including humans) and natural categories (including animals). With fMRI, we observed a separation of face and body categories from all others. As noted above, this difference with the prior literature could reflect a broader sampling of categories in our study or the use of backgrounds rather than segmented objects presented in isolation (Kriegeskorte et al., 2008; Sha et al., 2015). However, evidence for an animate distinction has been reported even with a large sampling of natural scenes (Naselaris et al., 2012). Alternatively, it is also possible that what has been termed animacy in previous studies primarily reflects the presence of face or body features and not animacy *per se*. Indeed, a recent study found that animate objects (e.g. cow) and inanimate objects that looked like an animate object (e.g. cow-shaped mug) are represented similary in vTC (Bracci et al., 2017b).

5. Conclusion

By comparing behavioral similarity judgments with fMRI responses in visual cortex across a range of object and scene categories, we find that while there is a correlation between fMRI and behavior, particularly in scene-selective areas, the structure of representations is strikingly different. While there appears to be no simple rigid transformation that might bring the behavioral and brain representational spaces into alignment, investigating the nature of the transformation necessary to align these spaces is a key question for future research. Further, while both the behavior and the fMRI data correlate well with DNN features, the modalities best matched different levels of representation. Collectively, these results suggest i) that there is only a limited correspondence between localized fMRI responses and behavioral similarity judgments with each domain capturing different properties of the images, and ii) that in comparing representational spaces, it is not sufficient to focus only on correlations between those spaces.

Conflicts of interest

None.

Author contributions

MLK, DJK and CIB designed the study. MLK and IIAG performed the research. MLK, IIAG, AS and DJK analyzed the data. MLK, DJK, IIAG, AS and CIB wrote the paper.

Acknowledgements

We thank Susan Wardle and Martin Hebart for helpful discussion and comments on earlier versions of this manuscript, Ed Silson for help with the ROI definitions, and Steven Scholte for help implementing the DNN analyses. This research was supported by the Intramural Research Program of the US National Institute of Mental Health (ZIAMH 002909), Clinical Study Protocol 93-M-0170, NCT00001360. MLK was also supported by the National Institutes of Health Predoctoral Training Grant (T32-GM108540). The authors declare no competing financial interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.04.079.

References

Bankson, B.B., Hebart, M.N., Groen, I.I.A., Baker, C.I., 2018. The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. Neuroimage 178, 172–182.

Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A., 2017. Network dissection: quantifying interpretability of deep visual representations. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, p. 3319.

Bauer, A.J., Just, M.A., 2019. Brain reading and behavioral methods provide complementary perspectives on the representation of concepts. Neuroimage 186, 794–805.

Bonner, M.F., Epstein, R.A., 2018. Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. PLoS Comput. Biol. 14 https://doi.org/10.1371/journal.pcbi.1006111 e1006111.

Bracci, S., Daniels, N., Op de Beeck, H., 2017a. Task context overrules object- and category-related representational content in the human parietal cortex. Cerebr. Cortex 27, 310–321.

Bracci, S., Kalfas, I., Op de Beeck, H., 2017b. The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. BioRxiv. https://doi.org/10.1101/228932.

Carlson, T.A., Simmons, R.A., Kriegeskorte, N., Slevc, L.R., 2014. The emergence of semantic meaning in the ventral temporal pathway. J. Cogn. Neurosci. 26, 120–131.

Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Return of the Devil in the Details: Delving Deep into Convolutional Nets (arXiv).

Cichy, R.M., 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. Sci. Rep. 6, 27755.

Cichy, R.M., Kriegeskorte, N., Jozwik, K.M., van den Bosch, J.J.F., Charest, I., 2019. The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. Neuroimage 194, 12–24. https://doi.org/10.1016/j.neuroimage.2019.03.031.

Clarke, A., Devereux, B.J., Randall, B., Tyler, L.K., 2015. Predicting the time course of individual objects with MEG. Cerebr. Cortex 25, 3602–3612.

Clarke, A., Tyler, L.K., 2014. Object-specific semantic coding in human perirhinal cortex. J. Neurosci. 34, 4766–4775.

Connolly, A.C., Guntupalli, J.S., Gors, J., Hanke, M., Halchenko, Y.O., Wu, Y.-C., Abdi, H., Haxby, J.V., 2012. The representation of biological classes in the human brain. J. Neurosci. 32, 2608–2618.

Çukur, T., Huth, A.G., Nishimoto, S., Gallant, J.L., 2013. Functional subdomains within human FFA. J. Neurosci. 33, 16748–16766.

Çukur, T., Huth, A.G., Nishimoto, S., Gallant, J.L., 2016. Functional subdomains within scene-selective cortex: parahippocampal place area, retrosplenial complex, and occipital place area. J. Neurosci. 36, 10257–10273.

Devereux, B.J., Clarke, A., Tyler, L.K., 2018. Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. Sci. Rep. 8, 10636. https://doi.org/10.1038/s41598-018-28865-1.

Epstein, R.A., 2008. Parahippocampal and retrosplenial contributions to human spatial navigation. Trends Cogn. Sci. (Regul Ed) 12, 388–396.

Erez, Y., Duncan, J., 2015. Discrimination of visual categories based on behavioral relevance in widespread regions of frontoparietal cortex. J. Neurosci. 35, 12383–12393.

Freeman, J., Brouwer, G.J., Heeger, D.J., Merriam, E.P., 2011. Orientation decoding depends on maps, not columns. J. Neurosci. 31, 4792–4804.
Greene, M.R., Baldassano, C., Esteva, A., Beck, D.M., Fei-Fei, L., 2016. Visual scenes are categorized by function. J. Exp. Psychol. Gen. 145, 82–94.
Grill-Spector, K., Weiner, K.S., 2014. The functional architecture of the ventral temporal cortex and its role in categorization. Nat. Rev. Neurosci. 15, 536–548.
Groen II, Greene, M.R., Baldassano, C., Fei-Fei, L., Beck, D.M., Baker, C.I., 2018. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. Elife 7.
Groen II, Silson, E.H., Baker, C.I., 2017. Contributions of low-and high-level properties to neural processing of visual scenes in the human brain. Phil Trans R Soc B 372.
Güçlü, U., van Gerven, M.A.J., 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J. Neurosci. 35, 10005–10014.
Güçlütürk, Y., Güçlü, U., van Gerven, M., van Lier, R., 2018. Representations of naturalistic stimulus complexity in early and associative visual and auditory cortices. Sci. Rep. 8, 3439.
Guntupalli, J.S., Hanke, M., Halchenko, Y.O., Connolly, A.C., Ramadge, P.J., Haxby, J.V., 2016. A model of representational spaces in human cortex. Cerebr. Cortex 26, 2919–2934.
Harel, A., Bentin, S., 2009. Stimulus type, level of categorization, and spatial-frequencies utilization: implications for perceptual categorization hierarchies. J. Exp. Psychol. Hum. Percept. Perform. 35, 1264–1273.
Harel, A., Kravitz, D.J., Baker, C.I., 2014. Task context impacts visual object processing differentially across the cortex. Proc. Natl. Acad. Sci. U.S.A. 111, E962–E971.
Hasson, U., Harel, M., Levy, I., Malach, R., 2003. Large-scale mirror-symmetry organization of human occipito-temporal object areas. Neuron 37, 1027–1041.
Hasson, U., Levy, I., Behrmann, M., Hendler, T., Malach, R., 2002. Eccentricity bias as an organizing principle for human high-order object areas. Neuron 34, 479–490.
Haxby, J.V., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., Hanke, M., Ramadge, P.J., 2011. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72, 404–416.
Hebart, M.N., Bankson, B.B., Harel, A., Baker, C.I., Cichy, R.M., 2018. The representational dynamics of task and object processing in humans. Elife 7.
Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L., 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76, 1210–1224.
Iordan, M.C., Greene, M.R., Beck, D.M., Fei-Fei, L., 2015. Basic level category structure emerges gradually across human ventral visual cortex. J. Cogn. Neurosci. 27, 1427–1446.
Iordan, M.C., Greene, M.R., Beck, D.M., Fei-Fei, L., 2016. Typicality sharpens category representations in object-selective cortex. Neuroimage 134, 170–179.
Kanwisher, N., 2010. Functional specificity in the human brain: a window into the functional architecture of the mind. Proc. Natl. Acad. Sci. U.S.A. 107, 11163–11170.
Kanwisher, N., Dilks, D.D., 2013. The functional organization of the ventral visual pathway in humans. New Visual Neurosci. 733–748.
Khaligh-Razavi, S.-M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol. 10, e1003915.
Konkle, T., Oliva, A., 2012. A real-world size organization of object responses in occipitotemporal cortex. Neuron 74, 1114–1124.
Kravitz, D.J., Kriegeskorte, N., Baker, C.I., 2010. High-level visual object representations are constrained by position. Cerebr. Cortex 20, 2916–2925.
Kravitz, D.J., Peng, C.S., Baker, C.I., 2011. Real-world scene representations in high-level visual cortex: it's the spaces more than the places. J. Neurosci. 31, 7322–7333.
Kravitz, D.J., Saleem, K.S., Baker, C.I., Ungerleider, L.G., Mishkin, M., 2013. The ventral visual pathway: an expanded neural framework for the processing of object quality. Trends Cogn. Sci. (Regul Ed) 17, 26–49.
Kriegeskorte, N., 2015. Deep neural networks: a new framework for modeling biological vision and brain information processing. Annu. Rev. Vis. Sci. 1, 417–446.
Kriegeskorte, N., Mur, M., 2012. Inverse MDS: inferring dissimilarity structure from multiple item arrangements. Front. Psychol. 3, 245.
Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., Bandettini, P.A., 2008. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60, 1126–1141.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 1097.
Larsson, J., Heeger, D.J., 2006. Two retinotopic visual areas in human lateral occipital cortex. J. Neurosci. 26, 13128–13142.
Malcolm, G.L., Groen, I.I.A., Baker, C.I., 2016. Making sense of real-world scenes. Trends Cogn. Sci. (Regul Ed) 20, 843–856.
Martin, A., 2016. GRAPES-Grounding representations in action, perception, and emotion systems: how object properties and categories are represented in the human brain. Psychon. Bull. Rev. 23, 979–990.
Martin, C.B., Douglas, D., Newsome, R.N., Man, L.L., Barense, M.D., 2018. Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. Elife 7.
Martin Cichy, R., 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.
Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P.A., Kriegeskorte, N., 2013. Human object-similarity judgments reflect and transcend the primate-IT object representation. Front. Psychol. 4, 128.
Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. Neuroimage 56, 400–410.
Naselaris, T., Stansbury, D.E., Gallant, J.L., 2012. Cortical representation of animate and inanimate objects in complex natural scenes. J. Physiol. Paris 106, 239–249.
Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A toolbox for representational similarity analysis. PLoS Comput. Biol. 10, e1003553.
Oosterhof, N.N., Connolly, A.C., Haxby, J.V., 2016. CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU octave. Front. Neuroinf. 10, 27.
Patterson, K., Nestor, P.J., Rogers, T.T., 2007. Where do you know what you know? The representation of semantic knowledge in the human brain. Nat. Rev. Neurosci. 8, 976–987.
Peelen, M.V., Downing, P.E., 2017. Category selectivity in human visual cortex: beyond visual object recognition. Neuropsychologia 105, 177–183.
Proklova, D., Kaiser, D., Peelen, M.V., 2016. Disentangling representations of object shape and object category in human visual cortex: the animate-inanimate distinction. J. Cogn. Neurosci. 28, 680–692.
Scholte, H.S., 2018. Fantastic DNimals and where to find them. Neuroimage 180, 112–113. https://doi.org/10.1016/j.neuroimage.2017.12.077.
Schyns, P.G., Oliva, A., 1999. Dr. Angry and Mr. Smile: when categorization flexibly modifies the perception of faces in rapid visual presentations. Cognition 69, 243–265.
Sha, L., Haxby, J.V., Abdi, H., Guntupalli, J.S., Oosterhof, N.N., Halchenko, Y.O., Connolly, A.C., 2015. The animacy continuum in the human ventral vision pathway. J. Cogn. Neurosci. 27, 665–678.
Silson, E.H., Chan, A.W.-Y., Reynolds, R.C., Kravitz, D.J., Baker, C.I., 2015. A retinotopic basis for the division of high-level scene processing between lateral and ventral human occipitotemporal cortex. J. Neurosci. 35, 11921–11935.
Silson, E.H., Groen, I.I.A., Kravitz, D.J., Baker, C.I., 2016a. Evaluating the correspondence between face-, scene-, and object-selectivity and retinotopic organization within lateral occipitotemporal cortex. J. Vis. 16, 14.
Silson, E.H., Steel, A.D., Baker, C.I., 2016b. Scene-selectivity and retinotopy in medial parietal cortex. Front. Hum. Neurosci. 10, 412. https://doi.org/10.3389/fnhum.2016.00412.
Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44, 83–98.
Tripp, B., 2017. A deeper understanding of the brain. Neuroimage 180, 114–116. https://doi.org/10.1016/j.neuroimage.2017.12.079.
Van de Moortele, P.-F., Auerbach, E.J., Olman, C., Yacoub, E., Uğurbil, K., Moeller, S., 2009. T1 weighted brain images at 7 Tesla unbiased for Proton Density, T2* contrast and RF coil receive B1 sensitivity with simultaneous vessel visualization. Neuroimage 46, 432–446.
Van Uden, C.E., Nastase, S.A., Connolly, A.C., Feilong, M., Hansen, I., Gobbini, M.I., Haxby, J.V., 2018. Modeling semantic encoding in a common neural representational space. Front. Neurosci. 12, 437.
Vaziri-Pashkam, M., Xu, Y., 2017. Goal-directed visual processing differentially impacts human ventral and dorsal visual representations. J. Neurosci. 37, 8767–8782.
Vedaldi, A., Lenc, K., 2015. Matconvnet: convolutional neural networks for MATLAB. In: Proceedings of the 23rd ACM International Conference on Multimedia - MM '15. ACM Press, New York, New York, USA, pp. 689–692.
Watson, D.M., Andrews, T.J., Hartley, T., 2017. A data driven approach to understanding the organization of high-level visual cortex. Sci. Rep. 7, 3596.
Wen, H., Shi, J., Chen, W., Liu, Z., 2018. Deep residual network predicts cortical representation and organization of visual features for rapid categorization. Sci. Rep. 8, 3752.
Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., Liu, Z., 2017. Neural encoding and decoding with deep learning for dynamic natural vision. Cerebr. Cortex 28, 1–25.
Yamins, D.L.K., DiCarlo, J.J., 2016. Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. 19, 356–365.
Yamins, D.L.K., 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. Proc. Natl. Acad. Sci. U.S.A. 111, 8619–8624.
Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision, p. 818.
Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A., 2014. Learning deep features for scene recognition using places database. Adv. Neural Inf. Process. Syst. 487.