Examining the Coding Strength of Object Identity and Nonidentity Features in Human Occipito-Temporal Cortex and Convolutional Neural Networks

Yaoda Xu¹ and Maryam Vaziri-Pashkam²

¹Department of Psychology, Yale University, New Haven, Connecticut 06520, and ²National Institute of Mental Health, Bethesda, Maryland 20892-9663

A visual object is characterized by multiple visual features, including its identity, position and size. Despite the usefulness of identity and nonidentity features in vision and their joint coding throughout the primate ventral visual processing pathway, they have so far been studied relatively independently. Here in both female and male human participants, the coding of identity and nonidentity features was examined together across the human ventral visual pathway. The nonidentity features tested included two Euclidean features (position and size) and two non-Euclidean features (image statistics and spatial frequency (SF) content of an image). Overall, identity representation increased and nonidentity feature representation decreased along the ventral visual pathway, with identity outweighing the non-Euclidean but not the Euclidean features at higher levels of visual processing. In 14 convolutional neural networks (CNNs) pretrained for object categorization with varying architecture, depth, and with/without recurrent processing, nonidentity feature representation showed an initial large increase from early to mid-stage of processing, followed by a decrease at later stages of processing, different from brain responses. Additionally, from lower to higher levels of visual processing, position became more underrepresented and image statistics and SF became more overrepresented compared with identity in CNNs than in the human brain. Similar results were obtained in a CNN trained with stylized images that emphasized shape representations. Overall, by measuring the coding strength of object identity and nonidentity features together, our approach provides a new tool for characterizing feature coding in the human brain and the correspondence between the brain and CNNs.

Key words: features; fMRI; invariance; visual object representation

Significance Statement

This study examined the coding strength of object identity and four types of nonidentity features along the human ventral visual processing pathway and compared brain responses with those of 14 convolutional neural networks (CNNs) pretrained to perform object categorization. Overall, identity representation increased and nonidentity feature representation decreased along the ventral visual pathway, with some notable differences among the different nonidentity features. CNNs differed from the brain in a number of aspects in their representations of identity and nonidentity features over the course of visual processing. Our approach provides a new tool for characterizing feature coding in the human brain and the correspondence between the brain and CNNs.

Introduction

A major research focus in the past decades has been on understanding how the primate visual system recognizes an object regardless of how it may appear in the real world. In fact, discounting nonidentity features, such as position and size, and forming identity-preserving transformation-tolerant object representations has been regarded as the defining feature of primate high-level vision (DiCarlo and Cox, 2007; DiCarlo et al., 2012).

Vision, however, is not just about object recognition; it also helps us to interact with the objects in the world. For example, to pick up an object, its position, size, and orientation, rather than its identity, would be most relevant. In fact, both object identity and nonidentity features are represented together in higher

Received July 30, 2020; revised Mar. 12, 2021; accepted Mar. 15, 2021.

Author contributions: Y.X. designed research; Y.X. and M.V.-P. performed research; Y.X. analyzed data; Y.X. wrote the paper.

The project is supported by National Institutes of Health (NIH) Grants 1R01-EY-022355 and 1R01-EY-030854 to Y.X. M.V.-P. was supported in part by NIH Intramural Research Program Grant ZIA MH002035. We thank Martin Schrimpf for help implementing CORnet-S; JohnMark Tayler for extracting the features from the three Resnet-50 models trained with the stylized images; and Thomas O'Connell, Brian Scholl, JohnMark Taylor, and Nick Turk-Brown for helpful discussions and feedback on the results.

The authors declare no competing financial interests.

Correspondence should be addressed to Yaoda Xu at xucogneuro@gmail.com.

<https://doi.org/10.1523/JNEUROSCI.1993-20.2021>

Copyright © 2021 the authors

visual areas in occipito-temporal cortex (OTC) in both macaques (Hung et al., 2005; Hong et al., 2016) and humans (Eger et al., 2008; Konen and Kastner, 2008; Kravitz et al., 2008, 2010; Sayres and Grill-Spector, 2008; Schwarzklose et al., 2008; Carlson et al., 2011; Cichy et al., 2011, 2013; Reithler et al., 2017; Vaziri-Pashkam et al., 2019); with these two types of information being represented in a largely distributed and overlapping manner (Hong et al., 2016). Despite their usefulness in vision and their joint coding in visual processing, object identity and nonidentity features have so far been studied relatively independently. Little is known regarding the relative coding strength for these two types of information within a brain region and how this relative strength changes along the ventral visual processing pathway.

Recent hierarchical convolutional neural networks (CNNs) have achieved human-like object categorization performance (Kriegeskorte, 2015; Yamins and DiCarlo, 2016; Rajalingham et al., 2018; Serre, 2019), with representations formed in early and late layers of the network tracking those of the earlier and later human visual processing regions, respectively (Khaligh-Razavi and Kriegeskorte, 2014; Cichy et al., 2016; Eickenberg et al., 2017; Güçlü and van Gerven, 2017). Together with results from monkey neurophysiological studies, CNNs have been regarded by some as the current best models of the primate visual system (Cichy and Kaiser, 2019; Kubilius et al., 2019). Nevertheless, we lack a detailed and clear understanding of how information is processed in CNNs. This is especially evident from recent studies reporting a number of discrepancies between the brain and CNNs (Serre, 2019). Our own investigation has shown that the close brain–CNN correspondence was rather limited and could not fully capture visual processing in the human brain (Xu and Vaziri-Pashkam, 2021). By examining the representations of object identity and nonidentity features in CNNs and comparing them with those from the human brain, we may form a deeper understanding of visual processing in CNNs.

In this study, we analyzed four existing fMRI datasets and documented the coding strengths of object identity and nonidentity features both on their own and with respect to each other within a given brain region as well as across the human ventral visual processing pathway in OTC (Fig. 1). We examined four types of nonidentity features, including two Euclidean features (position and size) and two non-Euclidean features [image statistics and the spatial frequency (SF) content of an image]. We found an overall increase of identity and a decrease of nonidentity information representation along the ventral pathway. While identity representation dominated those of the non-Euclidean features examined here in higher levels of visual processing, this was not the case for the Euclidean features. We additionally examined 14 different CNNs pretrained to perform object categorization with varying architecture and depth, and with/without recurrent processing. Unlike the brain, nonidentity feature representation in CNNs showed an initial large increase followed by a decrease at later stages of processing. Additionally, from lower to higher levels of visual processing, position became more underrepresented, and image statistics and SF became more overrepresented compared with identity in CNNs than in the human brain.

Materials and Methods

fMRI experimental details. Details of the fMRI experiments have been described in two previously published studies (Vaziri-Pashkam and Xu, 2019; Vaziri-Pashkam et al., 2019). They are summarized here for the readers' convenience.

Seven (four females), 7 (four females), 6 (three females), and 10 (five females) healthy human participants with normal or corrected-to-normal visual acuity, all right handed, and between 18 and 35 years of age took part in experiments 1–4, respectively. All participants gave their written informed consent before the experiments and received payment for their participation. The experiments were approved by the Committee on the Use of Human Subjects at Harvard University. Each main experiment was performed in a separate session lasting between 1.5 and 2 h. Each participant also completed two additional sessions for topographic mapping and functional localizers. MRI data were collected using a MAGNETOM Trio, A Tim System 3 T scanner (Siemens), with a 32-channel receiver array head coil. For all the fMRI scans, a T2*-weighted gradient echo pulse sequence with a TR of 2 s and voxel size of $3 \times 3 \times 3$ mm was used. fMRI data were analyzed using FreeSurfer (<https://surfer.nmr.mgh.harvard.edu>), FsFast (Dale et al., 1999), and in-house MATLAB codes. fMRI data preprocessing included 3D motion correction, slice timing correction, and linear and quadratic trend removal. Following standard practice, a general linear model was then applied to the fMRI data to extract β -weights as response estimates.

The general experimental paradigm consisted of a one-back image repetition detection task in which participants viewed a stream of sequentially presented images and pressed a response button when the same image repeated back to back (Fig. 1A). This task engaged participants' attention on the object shapes and ensured robust fMRI responses. Two image repetitions occurred randomly in each image block. We used cutout grayscale images from eight real-world object categories (faces, bodies, houses, cats, elephants, cars, chairs, and scissors) and modified them to occupy approximately the same area on the screen (Fig. 1B). For each object category, we selected 10 exemplar images that varied in identity, pose, and viewing angle to minimize the low-level similarities among them. Each block of image presentation contained images from the same object category. Participants fixated at a central red dot throughout the experiment. Eye movements were monitored in all the fMRI experiments to ensure proper fixation. We examined responses from early visual areas V1 to V4 and higher visual processing regions in lateral occipito-temporal (LOT) and ventral occipito-temporal (VOT) cortices (see more details below; Fig. 1C).

In experiment 1, we tested position tolerance and presented images either above or below the fixation (Fig. 1D). Each block contained a random sequential presentation of 10 exemplars from the same object category shown either all above or all below the fixation. To ensure that object identity representation in lower brain regions truly reflected the representation of object identity and not low-level differences among the images of the different categories, controlled images with low-level image differences equated among the different categories were shown. These controlled images were generated by equalizing contrast, luminance, and spatial frequency of the images across all the categories using the SHINE toolbox (Willenbockel et al., 2010; Fig. 1D). All images subtended $2.9^\circ \times 2.9^\circ$ and were shown at 1.56° above the fixation in half of the 16 blocks and the same distance below the fixation in the other half of the blocks. Each image was presented for 200 ms followed by a 600 ms blank interval between the images. Each experimental run contained 16 blocks, 1 for each of the 8 categories in each of the two image positions. The order of the eight object categories and the two positions were counterbalanced across runs and participants. Each block lasted 8 s and was followed by an 8 s fixation period. There was an additional 8 s fixation period at the beginning of the run. Each participant completed one scan session with 16 runs for this experiment, each lasting 4 min and 24 s.

In experiment 2, we tested size tolerance and presented images either in a large size ($5.77^\circ \times 5.77^\circ$) or small size ($2.31^\circ \times 2.31^\circ$) centered at fixation (Fig. 1D). As in experiment 1, controlled images were used here. Half of the 16 blocks contained small images, and the other half, large images. Other details of the experiment were identical to those of experiment 1.

In experiment 3, we tested image statistics tolerance and presented images at fixation either in the original unaltered format or in the controlled format (subtended $4.6^\circ \times 4.6^\circ$; Fig. 1D). As mentioned in stimulus description of experiment 1, the controlled images were generated by

Figure 1. *A*, An illustration of the block design paradigm used. Participants performed a one-back repetition detection task on the images. An actual block in the experiment contained 10 images with two repetitions per block. See Materials and Methods for more details. *B*, The eight real-world object categories used. *C*, The brain regions examined. They included topographically defined early visual areas V1 to V4 and functionally defined higher object-processing regions LOT and VOT. *D*, The four types of nonidentity transformations examined. They included two Euclidean transformations—position and size—and two non-Euclidean transformations—image statistics and spatial frequency. *E*, The calculation of Euclidean distance and identity dominance indices without noise correction. To quantify identity representation from the averaged z-normalized pattern vectors, the Euclidean distance between vectors of two different object categories sharing the same value of a nonidentity feature is calculated as d_{between} . To quantify nonidentity feature representation, the Euclidean distance between vectors of the same object category having two different values of a nonidentity feature is calculated as d_{within} . To quantify the relative coding strength of identity and nonidentity features, an identity dominance index is computed as the difference between the two distance measures divided by the sum of the two measures. A lack of nonidentity feature representation would result in d_{within} of 0 and an index of 1; conversely, a lack of identity representation would result in d_{between} of 0 and an index of -1 . *F*, The calculation of Euclidean distance and identity dominance indices with noise correction. Because of fMRI measurement noise, even a lack of representation would result in a positive Euclidean distance measure. Consequentially, d_{between} and d_{within} may never be 0 and the index

equalizing the image contrast, luminance, and spatial frequency across all the categories using the SHINE toolbox (Willenbockel et al., 2010). Half of the 16 blocks contained original images, and the other half, controlled images. Other details of the experiment were identical to those of experiment 1.

In experiment 4, only six of the original eight object categories were included, and they were faces, bodies, houses, elephants, cars, and chairs. Images were shown in the following three conditions: Full-SF, High-SF, and Low-SF (Fig. 1D). In the Full-SF condition, the full-spectrum images were shown without modification of the SF content. In the High-SF condition, images were high-pass filtered using a finite impulse response (FIR) filter with a cutoff frequency of 4.40 cycles/°. In the Low-SF condition, the images were low-pass filtered using an FIR filter with a cutoff frequency of 0.62 cycles/°. The DC component was restored after filtering so that the image backgrounds were equal in luminance. Each run contained 18 blocks, 1 for each of the category and SF condition combinations. Each participant completed a single scan session containing 18 experimental runs, each lasting 5 min. Other details of the experimental design were identical to those of experiment 1. Only the results from the low-frequency and high-frequency conditions were included in the present analysis.

We examined responses from independent localized early visual areas V1 to V4 and higher visual processing regions LOT and VOT (Fig. 1C). V1 to V4 were mapped with flashing checkerboards using standard techniques (Serenio et al., 1995). Following the detailed procedures described in the study by Swisher et al. (2007) and by examining phase reversals in the polar angle maps, we identified areas V1 to V4 in the occipital cortex of each participant (see also Bettencourt and Xu, 2016; Fig. 1C). To identify LOT and VOT, following Kourtzi and Kanwisher (2000), participants viewed blocks of intact and scrambled object images. The object images contained everyday natural and manmade objects and did not include any objects that would evoke category-selective responses in the ventral visual cortex (e.g., faces, scenes, or bodies). LOT and VOT were defined as a cluster of continuous voxels in the lateral and ventral occipital cortex, respectively, that responded more to the intact than to the scrambled object images. LOT and VOT loosely correspond to the location of lateral occipital and posterior fusiform object selective regions (Malach et al., 1995; Grill-Spector et al., 1998; Kourtzi and Kanwisher, 2000) but extend farther into the temporal cortex in an effort to include as many object-selective voxels as possible in occipito-temporal regions.

To generate the fMRI response pattern for each ROI in a given run, we first convolved an 8 s stimulus presentation boxcar (corresponding to the length of each image block) with a hemodynamic response function to each condition; we then conducted a general linear model analysis to extract the β -weight for each condition in each voxel of that ROI. These voxel β -weights were used as the fMRI response pattern for that condition in that run. Following the study by Tarhan and Konkle (2020), to increase power, we selected the top 75 most reliable voxels in each ROI for further analyses. This was done by splitting the data into odd and even halves, averaging the data across the runs within each half, correlating the β -weights from all the conditions between the two halves for each voxel, and then selecting the top 75 voxels showing the highest correlation. This is akin to including the best units in monkey neurophysiological studies. For example, Cadieu et al. (2014) selected only a small subset of all recorded single units for their brain–CNN analysis. We

obtained the fMRI response pattern for each condition from the 75 most reliable voxels in each ROI of each run. Very similar results were obtained if we included all voxels in an ROI instead of just the 75 most reliable voxels (see Fig. 8A).

CNN details. We included 14 CNNs in our analyses (Table 1). They included both shallower networks, such as Alexnet, VGG16, and VGG 19, and deeper networks, such as Googlenet, Inception-v3, Resnet-50 (RN50), and Resnet-101. We also included a recurrent network, Cornet-S, that has been shown to capture the recurrent processing in macaque inferior temporal (IT) cortex with a shallower structure (Kar et al., 2019; Kubilius et al., 2019). This CNN has been recently argued to be the current best model of the primate ventral visual processing regions (Kar et al., 2019). All the CNNs used were pretrained with ImageNet images (Deng et al., 2009).

To understand how the specific training images would impact CNN representations, in addition to CNNs trained with ImageNet images, we also examined Resnet-50 trained with stylized ImageNet images (Geirhos et al., 2019). We examined the representations formed in Resnet-50 pretrained with three different procedures (Geirhos et al., 2019): trained only with the stylized ImageNet Images (RN50-SIN), trained with both the original and the stylized ImageNet Images (RN50-SININ), and trained with both sets of images and then fine-tuned with the stylized ImageNet images (RN50-SININ-IN).

Following the study by O’Connell and Chun (2018), we sampled between 6 and 11 mostly pooling and fully-connected (FC) layers of each CNN (Table 1, specific CNN layers sampled). Pooling layers were selected because they typically mark the end of processing for a block of layers before information is pooled and passed on to the next block of layers. When there were no obvious pooling layers present, the last layer of a block was chosen. For a given CNN layer, we extracted the CNN layer output for each object image in a given condition, averaged the output from all images in a given category for that condition, and then z-normalized the responses to generate the CNN layer response for that object category in that condition (similar to how fMRI category responses were extracted). Cornet-S and the different versions of Resnet-50 were implemented in Python. All other CNNs were implemented in MATLAB. Output from all CNNs was analyzed and compared with brain responses using MATLAB.

Quantifying the presence of object identity and nonidentity feature representation in human OTC and CNNs. To verify that the representation of object identity and nonidentity information was indeed present in the current dataset before we assess the relative coding strength of these features and to account for noise/reliability differences among the different brain regions, we used a split-half approach. Specifically, we first averaged the data within the odd and even halves of the runs separately. We then applied z-normalization to the averaged pattern of each condition in each ROI in each half of the runs separately to remove amplitude differences between conditions and ROIs. Using experiment 2, in which objects appeared in either the small or large size as an example, we obtained a between-object Euclidean distance measure (d_{between}) by calculating the average Euclidean distance between two different object categories across the two halves of the runs while holding size constant (e.g., between cars in the small size in the odd runs and chairs in the small size in the even runs; Fig. 1F, illustration). We also obtained a same-object Euclidean distance measure (d_{same}) by calculating the average Euclidean distance between the same object category across the two halves of the runs while holding size constant (e.g., between cars in the small size in the odd runs and cars in the small size in the even runs; Fig. 1F, illustration). Because Euclidean distance measures were always positive, d_{same} allowed us to assess the amount of measurement noise in each brain region, with a larger amount of noise across odd and even runs resulting in a larger d_{same} . To account for noise differences across brain regions and to facilitate between-region comparisons, we corrected d_{between} in two different ways to compare how this would impact the results. Specifically, we either subtracted d_{same} from d_{between} or divided d_{between} by d_{same} to derive a subtraction-corrected distance measure, $d_{(s)} \text{ between}$, and a division-corrected distance measure $d_{(d)} \text{ between}$ for each stimulus condition (Fig. 1F).

←

would be compressed toward 0. To remedy this, d_{between} and d_{within} are calculated across the odd and even halves of the runs. The Euclidean distance between vectors of the same object category with the same value of a nonidentity feature across the odd and even halves of the runs is also calculated as d_{same} and provides a measure of the noise level in the data. d_{between} and d_{within} are then noise corrected by either subtracting or dividing by d_{same} to generate subtraction-corrected distance measures, $d_{(s)} \text{ between}$ and $d_{(s)} \text{ within}$, and division-corrected distance measures $d_{(d)} \text{ between}$ and $d_{(d)} \text{ within}$. This allows us to assess whether or not results may differ between these two correction methods (Figs. 2, 3). Noise-corrected identity dominance is computed using $d_{(s)} \text{ between}$ and $d_{(s)} \text{ within}$ by applying the same formula as in *E* (note that in the absence of measurement noise, the corrected and the uncorrected indices converge).

Table 1. The CNNs and the layers examined in this study

CNN name	Depth/blocks	Layers	Layers sampled, <i>n</i>	Sampled layer names and locations (indicated in the parenthesis)
Alexnet	8	25	6	"pool1" (5), "pool2" (9), "pool5" (16), "fc6" (17), "fc7" (20), "fc8" (23)
Cornet-S	4	42	6	"V1_output" (8), "V2_output" (18), "V4_output" (28), "IT_output" (38), "decoder_avgpool" (39), "decoder_output" (42)
Densenet-201	201	709	6	"pool1" (6), "pool2_pool" (52), "pool3_pool" (140), "pool4_pool" (480), "avg_pool" (706), "fc1000" (707)
Googlenet	22	144	6	"pool1-3×3_s2" (4), "pool2-3×3_s2" (11), "pool3-3×3_s2" (40), "pool4-3×3_s2" (111), "pool5-7×7_s1" (140), "loss3-classifier" (142)
Inception_v3	48	316	11	"average_pooling2d_1" (29), "average_pooling2d_2" (52), "average_pooling2d_3" (75), "average_pooling2d_4" (121), "average_pooling2d_5" (153), "average_pooling2d_6" (185), "average_pooling2d_7" (217), "average_pooling2d_8" (264), "average_pooling2d_9" (295), "avg_pool" (313), "predictions" (314)
Inception-resnet_v2	164	825	7	"max_pooling2d_1" (12), "max_pooling2d_2" (19), "average_pooling2d_1" (29), "max_pooling2d_3" (285), "max_pooling2d_4" (648), "avg_pool" (822), "predictions" (823)
Mobilenet_v2	54	155	10	"block_2_project_BN" (26), "block_4_project_BN" (43), "block_6_project_BN" (61), "block_8_project_BN" (78), "block_10_project_BN" (96), "block_12_project_BN" (113), "block_14_project_BN" (130), "block_16_project_BN" (148), "global_average_pooling2d_1" (152), "Logits" (153)
Resnet-18	18	72	6	"pool1" (6), "res2b_relu" (20), "res3b_relu" (36), "res4b_relu" (52), "pool5" (69), "fc1000" (70)
Resnet-50	50	177	6	"max_pooling2d_1" (5), "activation_10_relu" (37), "activation_22_relu" (79), "activation_40_relu" (141), "avg_pool" (174), "fc1000" (175)
Resnet-101	101	347	6	"pool1" (5), "res2c_relu" (37), "res3b3_relu" (79), "res4b22_relu" (311), "pool5" (344), "fc1000" (345)
SqueezeNet	18	68	5	"pool1" (4), "pool3" (19), "pool5" (34), "conv10" (64), "pool10" (66)
Vgg-16	16	41	8	"pool1" (6), "pool2" (11), "pool3" (18), "pool4" (25), "pool5" (32), "fc6" (33), "fc7" (36), "fc8" (39)
Vgg-19	19	47	8	"pool1" (6), "pool2" (11), "pool3" (20), "pool4" (29), "pool5" (38), "fc6" (39), "fc7" (42), "fc8" (45)
Xception	71	171	8	"block2_pool" (18), "block4_pool" (42), "block6_sepconv3_bn" (68), "block8_sepconv3_bn" (94), "block10_sepconv3_bn" (120), "block12_sepconv3_bn" (146), "avg_pool" (168), "predictions" (169)

Following O'Connell and Chun (2018), we sampled between 6 and 11 mostly pooling and FC layers of each CNN. Pooling layers were selected because they typically mark the end of processing for a block of layers before information is pooled and passed on to the next block of layers. When there were no obvious pooling layers present, the last layer of a block was chosen.

If a brain region contained distinctive representations for the different object categories, then the distance between two different categories between the odd and even runs should be greater than the distance between the same category across the odd and even runs. This would result in $d_{(s) \text{ between}}$ to be greater than zero and $d_{(d) \text{ between}}$ to be greater than one. This was assessed at the participant group level using one-tailed *t* tests.

With these corrected Euclidean distance measures, we also examined whether the strength of object representations showed a linearly increasing or decreasing trend across brain regions and the magnitude of such a trend. To do so, using Pearson correlation, we correlated each participant's corrected Euclidean distance measures across the six brain regions with the rank order of these brain regions (i.e., from 1 to 6). We then tested the resulting Fisher-transformed correlation coefficient against 0 at the group level using two-tailed *t* tests (Fisher transformation was applied here, and all subsequent statistical comparisons involving correlation coefficients to ensure normal distribution of the values). We additionally assessed differences in corrected Euclidean distances between lower and higher brain regions at the group level using two-tailed *t* tests. All *t* tests conducted were corrected for multiple comparisons using the method of Benjamini and Hochberg (1995).

To measure the coding strength of nonidentity feature representations in the brain, using the same split-half approach, we calculated within-object Euclidean distances (d_{within}), in which we measured the averaged distance between the same object category across the two values of a nonidentity feature (e.g., between cars in the small and large sizes; Fig. 1F, illustration) across odd and even runs and then corrected the distance by either subtracting or dividing d_{same} to obtain the corrected d_{within} using either the subtraction method ($d_{(s) \text{ within}}$) or the division method ($d_{(d) \text{ within}}$), as we did earlier (Fig. 1F). We repeated the same set of analyses as described above to document the representations of nonidentity features in the brain.

For the 14 CNNs examined, we also calculated d_{between} and d_{within} for each sampled layer. Given the absence of measurement noise in CNNs, these distance measures were computed directly from the category responses from each layer without a split-half approach (Fig. 1E, illustration). To facilitate between-layer comparisons, these distance measures were corrected for the number of units in each layer. This was done by dividing each distance measure by two times the square root of the total number of units in that layer—this was necessary as the distance between

two opposite patterns would increase with an increasing number of units/dimensions. For these normalized distance measures, we then computed the magnitude of a linear trend across the layers for each CNN as we did with the brain data. To provide a quantitative comparison between the brain and the CNNs, using two-tailed *t* test, we examined whether or not the Fisher-transformed correlation coefficient value obtained from a CNN was significantly different from the individual values of the human participants at the group level.

Visualizing and quantifying the relative coding strength of identity and nonidentity features. To directly visualize how object identity and nonidentity features may be represented together in a brain region, from the *z*-normalized fMRI response patterns averaged over all the runs, we first calculated all pairwise Euclidean distances, including all the categories shown and both values of each nonidentity feature (e.g., small and large). We then constructed a category representational dissimilarity matrix (RDM) for each brain region (Kriegeskorte and Kievit, 2013). Using multidimensional scaling (MDS; Shepard, 1980), we visualized this RDM by placing it on a 2D space with the distances among the different feature combinations approximating their similarities in representation (see Figs. 4, 5, 6, 7). We decided to present results from the average of all the runs here rather than the corrected split-half results because of its simplicity and because of the fact that correction would largely serve as a scaling factor and would not drastically change the approximate depictions of the identity and nonidentity features within a given brain region in the MDS plot (it is actually unclear how we may construct an RDM with the corrected Euclidean distance measures as the diagonal of the RDM would not be 0 because of how correction was applied). Since the MDS plots provide approximate, rather than precise, depictions of feature representations, and the precise representations were quantified by our index measure, as explained below, these MDS plots were provided as approximate visual illustrations of the actual representational space.

To quantify the relative coding strength of the different features in a brain region, using $d_{(s) \text{ within}}$ and $d_{(s) \text{ between}}$, we constructed an identity dominance index as follows:

$$\text{Identity dominance} = \left(d_{(s) \text{ between}} - d_{(s) \text{ within}} \right) / \left(d_{(s) \text{ between}} + d_{(s) \text{ within}} \right).$$

An identity-only representation that disregards the nonidentity feature would have a $d_{(s) \text{ within}}$ of 0 and an identity dominance of 1.

Figure 2. *A, B*, The coding strengths of object identity and position (*A*) and object identity and size (*B*) in human OTC and CNNs. The between-object Euclidean distances and the within-object Euclidean distances measure the coding strength of object identity and nonidentity features, respectively. The distance measures for the brain regions are corrected by the reliability of each region in two different ways, either by subtracting the reliability measure in brain (*s*) or by dividing by the reliability measure in brain (*d*; see Materials and Methods; Fig. 1*F*). Very similar results are obtained for both measures. Linear regression lines for each distance measure are included as gray dotted lines.

Conversely, a nonidentity-only representation that disregards identity would have a $d_{(s)}$ between of 0 and an identity dominance of -1 . An identity dominance of 0 indicates equal representational strength of object identity and nonidentity features, such that an object category is equally similar to itself in the other value of the nonidentity feature as it is to the other categories sharing the same value of the nonidentity feature. Here we used the subtraction-corrected Euclidean distances instead of the division-corrected Euclidean distances for the following two reasons. First, both types of corrections produced very similar results (Figs. 2, 3, and Results). Second, because of the absence of measurement noise, Euclidean distance measures from CNNs could be 0 when there is no representation of a feature. Using the subtraction-corrected Euclidean distance measures from the human brain would thus better match with the measures obtained from CNNs, which use the uncorrected Euclidean distances to calculate identity dominance (Fig. 1*E*, illustration). In fact, in the absence of measurement noise, the corrected and the uncorrected identity dominance would converge (Fig. 1, compare, *E* and *F*).

We quantified differences in identity dominance across brain regions using repeated-measures ANOVA. To examine whether identity dominance showed a linear increasing or decreasing trend across brain regions and the magnitude of such a trend, using Pearson correlation, we correlated each participant's identity dominance for the different brain regions with the rank order of these brain regions. We then tested the resulting Fisher-transformed correlation coefficient against 0 at the group level using two-tailed *t* tests. We additionally assessed differences between lower and higher brain regions using two-tailed *t* tests. As before, all *t* tests conducted were corrected for multiple comparisons using the method of Benjamini and Hochberg (1995).

To assess whether or not identity dominance would be consistent across the odd and even halves of the runs, when all runs were included and all voxels were included, we also derived identity dominance indices from the uncorrected Euclidean distances, as illustrated in Figure 1*E*. Note that because of the presence of measurement noise, the uncorrected Euclidean distances could never reach 0 even when there was a

Figure 3. *A, B*, The coding strengths of object identity and image statistics (*A*) and object identity and SF (*B*) in human OTC and CNNs. The between-object Euclidean distances and the within-object Euclidean distances measure the coding strength of object identity and nonidentity features, respectively. The distance measures for the brain regions are corrected by the reliability of each region used in two different ways, either by subtracting the reliability measure in brain (*s*) or by dividing by the reliability measure in brain (*d*; see Materials and Methods; Fig. 1*F*). Very similar results are obtained for both measures. Linear regression lines for each distance measure are included as gray dotted lines.

lack of feature representation; consequently, the identity dominance indices derived would be distorted and compressed toward 0 (i.e., never reaching 1 or -1). The uncorrected Euclidean distances were used for identity dominance calculation here because our reliability correction method involved pooling odd and even halves of the data together (Fig. 1*F*, illustration) and could not be directly applied to just the odd or just the even half of the data (we could in theory further divide the data into two halves within the odd or even runs to correct for reliability, but doing so would significantly lower power as fewer runs would be included). Since our goal was to assess whether or not data were consistent across the odd and even halves of the runs, when all runs were included, and when all voxels were included, deriving identity dominance indices from the uncorrected Euclidean distances provided an easy way to address this question.

For the CNN layer output, as illustrated in Figure 1*E*, because of the absence of measurement noise, we used the uncorrected Euclidean

distances and the same formula to calculate identity dominance indices. As with the brain data, we also computed the magnitude of a linear trend across the layers for each CNN. To provide a quantitative comparison between the brain and the CNNs, using two-tailed *t* test, we examined whether or not the Fisher-transformed correlation coefficient value obtained from a CNN was significantly different from the individual values of the human participants at the group level.

Experimental design and statistical analysis. The sample size for each fMRI experiment was chosen based on prior published studies (Haxby et al., 2001; Kamitani and Tong, 2005). Experimental design details of each individual fMRI experiment, details of the brain–CNN comparisons made (including the factors for each statistical test, the precise statistical tests, and planned comparisons made), and all critical details required for independent replication have been reported in detail in earlier parts of Materials and Methods. Correction for multiple comparisons has been applied using the method of Benjamini and Hochberg (1995)

whenever it was needed. All statistical tests were conducted in MATLAB.

Data availability. Data supporting the findings of this study are available at <https://osf.io/tsz47/>.

Results

Any given visual object input is characterized by multiple visual features, such as its identity, position, and size. Although object identity representation has been intensely studied in the primate brain, nonidentity features, such as position and size, have also been found to be robustly coded throughout the primate ventral visual processing pathway (Hong et al., 2016; Vaziri-Pashkam et al., 2019). Here we documented the coding strength of object identity and nonidentity features together in a brain region and how this would change across the ventral visual processing pathways in human OTC. We also examined responses from 14 CNNs pretrained to perform object categorization with varying architecture, depth, and the presence/absence of recurrent processing. Our fMRI data were collected with a block design in which responses were averaged over a whole block of multiple exemplars from the same category to increase signal to noise ratio (SNR) (Fig. 1A). A total of eight real-world object categories were included, and they were bodies, cars, cats, chairs, elephants, faces, houses, and scissors (Vaziri-Pashkam and Xu, 2019; Vaziri-Pashkam et al., 2019; Fig. 1B). The images were shown in conjunction with four nonidentity features (Fig. 1D), including two Euclidean features, position (top vs bottom) and size (small vs large), and two non-Euclidean features, image statistics (original vs controlled) and SF of an image (high SF vs low SF). Controlled images were created to achieve spectrum, histogram, and intensity normalization and equalization across the images from the different categories (Willenbockel et al., 2010). Controlled images also appeared in the position and size manipulations to ensure that object identity representation in lower brain regions would reflect the representation of object identity and not low-level differences among the images of the different categories.

To increase power, we extracted the averaged neural response patterns from each block of trials from the 75 most reliable voxels from each independently defined visual regions along the human OTC (see Materials and Methods; the same results were obtained when all voxels were included from each region; see Fig. 8A). These regions included topographic early visual areas V1 to V4 and higher visual object processing regions LOT and VOT (Fig. 1C). LOT and VOT have been considered as the homolog of the macaque IT cortex involved in visual object processing (Orban et al., 2004). Their responses have been shown to correlate with successful visual object detection and identification (Grill-Spector et al., 2000; Williams et al., 2007), and their lesions have been linked to visual object agnosia (Goodale et al., 1991; Farah, 2004).

The 14 CNNs we examined here included both shallower networks, such as Alexnet, VGG16, and VGG 19, and deeper networks, such as Googlenet, Inception-v3, Resnet-50, and Resnet-101 (Table 1). We also included a recurrent network, Cornet-S, that has been shown to capture the recurrent processing in macaque IT cortex with a shallower structure (Kar et al., 2019; Kubilius et al., 2019). This CNN has been recently argued to be the current best model of the primate ventral visual processing regions (Kar et al., 2019). To understand how the specific training images would impact CNN representations, in addition to CNNs trained with ImageNet images (Deng et al., 2009), we also examined Resnet-50 trained with stylized ImageNet images

(Geirhos et al., 2019). Following a previous study (O'Connell and Chun, 2018), we sampled from 6 to 11 mostly pooling layers of each CNN included (Table 1, specific CNN layers sampled). Pooling layers were selected because they typically mark the end of processing for a block of layers before information is passed on to the next block of layers.

We have previously compared the object representational structure similarity between the brain and the same 14 CNNs tested here in a subset of the data examined here (Xu and Vaziri-Pashkam, 2021). We found that lower visual areas showed better correlation with lower than higher CNN layers and the reverse was true for higher visual areas. This correspondence in brain–CNN representation was statistically significant in a majority of the CNNs examined here and replicated other published fMRI results (Khaligh-Razavi and Kriegeskorte, 2014; Cichy et al., 2016; Eickenberg et al., 2017; Güçlü and van Gerven, 2017). This provides a valid basis for conducting the present investigation in which we examined in further detail how identity and nonidentity features may be represented together in the brain and CNNs.

Quantifying the presence of object identity and nonidentity feature representation in human OTC and CNNs

Before we compare the relative coding strength of object identity and nonidentity features in human OTC, it is important to verify that the representation of such information is present in the current dataset. To test this and to account for noise/reliability differences among the different brain regions, we used a split-half approach. Specifically, we calculated the distance between different object categories and the distance between the same object category across the odd and even halves of the runs while holding the value of the nonidentity feature constant; we then either subtracted the latter from the former to compute a subtraction corrected between-object Euclidean distance ($d_{(s)} \text{ between}$) or divided the former by the latter to compute a division corrected between-object Euclidean distance ($d_{(d)} \text{ between}$; Fig. 1F, illustration; for more details, see Materials and Methods). Because both correction procedures are valid, we wanted to test how results might differ between these two procedures. The resulting distance measures for all four experiments are plotted in Figures 2 and 3. If a brain region contained distinctive representations for the different object categories, then when the value of the nonidentity feature was held constant, a greater distance would be expected between different than between the same object category across the odd and even runs. This would result in $d_{(s)} \text{ between} > 0$ and $d_{(d)} \text{ between} > 1$. This was found to be true for all brain regions examined in all four experiments (t values > 3.63 , p values < 0.01 , one-tailed, as only values greater than zero or one, respectively, were meaningful, and corrected for multiple comparisons using the method of Benjamini and Hochberg, 1995, for the six brain regions included in each experiment). This replicated our previously published results using a support vector machine (SVM) classifier to decode object responses in these brain regions (Vaziri-Pashkam et al., 2019; Vaziri-Pashkam and Xu, 2019).

With these corrected Euclidean distance measures, we additionally tested whether the strength of object identity representations showed a linearly increasing or decreasing trend across brain regions. To do so, using Pearson correlation, we correlated each participant's $d_{(s)} \text{ between}$ and $d_{(d)} \text{ between}$ values across the six brain regions with the rank order of these brain regions (i.e., from 1 to 6). We then tested the resulting Fisher-transformed correlation coefficient against 0 at the participant group level. All four experiments showed a positive trend (Table 2, the exact

Table 2. The correlation coefficients of between-object Euclidean distances and within-object Euclidean distances with the rank order of brain regions or CNN layers in the four experiments

System name	Between-object Euclidean distance				Within-object Euclidean distance			
	Position	Size	Image statistics	SF	Position	Size	Image statistics	SF
Brain - s (SE)	0.89 (0.03)	0.85 (0.03)	0.48 (0.20)	0.43 (0.15)	-0.88 (0.06)	-0.91 (0.03)	-0.87 (0.01)	-0.08 (0.17)
Brain - d (SE)	0.87 (0.04)	0.88 (0.02)	0.55 (0.25)	0.65 (0.11)	-0.80 (0.06)	-0.66 (0.11)	-0.90 (0.02)	0.10 (0.18)
Alexnet	0.47***/**	0.50***/**	0.68	-0.74***/**	-0.58**/*	-0.29**/*	0.61***/**	-0.01
Cornet-S	0.97**/**	0.98***/**	0.96**/*	0.61	-0.03***/**	0.64***/**	0.71***/**	0.89***/**
Densenet-201	0.92	0.96**/**	0.96**/*	0.95***/**	0.19***/**	0.89***/**	0.92***/**	0.87***/**
Googlenet	0.93	0.95**/*	0.92*†	-0.08**/**	-0.60**/*	0.13***/**	0.53***/**	0.43*
Inception_v3	0.92	0.96**/**	0.97**/*	0.95***/**	0.28***/**	0.77***/**	0.85***/**	0.88***/**
Inception-resnet_v2	0.88	0.89	0.88*	0.90**/**	-0.31***/**	0.85***/**	0.78***/**	0.82***/**
MobileNet_v2	0.88	0.89	0.92*†	0.81*	-0.06***/**	0.66***/**	0.59***/**	0.77***/**
Resnet-18	0.91	0.95**/*	0.94**/*	0.85**/*	0.20***/**	0.77***/**	0.84***/**	0.86***/**
Resnet-50	0.95††	0.97***/**	0.96**/*	0.71	0.23***/**	0.64***/**	0.68***/**	0.88***/**
Resnet-101	0.91	0.97***/**	0.96**/*	0.83*	0.08***/**	0.59***/**	0.62***/**	0.84***/**
SqueezeNet	0.75**/*	0.87	0.89*	0.21**	0.04***/**	0.58***/**	0.84***/**	0.84***/**
Vgg-16	0.51***/**	0.59**/**	0.68	-0.15***/**	-0.33***/**	0.17***/**	0.25***/**	0.00
Vgg-19	0.51***/**	0.62**/**	0.72	0.05**/**	-0.24***/**	0.19***/**	0.22***/**	0.24
Xception	0.81*†	0.88	0.81	0.68	-0.17***/**	0.51***/**	0.09***/**	0.56**/*†
Resnet-50-IN	0.95††	0.97***/**	0.96**/*	0.71	0.23***/**	0.64***/**	0.68***/**	0.88***/**
Resnet-50-SIN	0.95*†	0.96**/**	0.95**/*	0.93***/**	0.17***/**	0.73***/**	0.08***/**	0.89***/**
Resnet-50-SININ	0.91	0.96**/*	0.94**/*	0.87**/*	-0.08***/**	0.55***/**	0.47***/**	0.83***/**
Resnet-50-SININ-IN	0.91	0.96**/**	0.95**/*	0.88**/*	-0.01***/**	0.62***/**	0.65***/**	0.89***/**

A positive/negative correlation indicates that Euclidean distances increase/decrease over the course of processing. The between-object Euclidean distance and the within-object Euclidean distance reflected the coding strength of object identity and nonidentity features, respectively. The brain results are the average correlation coefficients across the group of human participants, with Euclidean distances corrected by the reliability of each brain region using either the subtraction (brain - s) or division (brain - d) method (see Materials and Methods) before correlations are computed. To provide some quantification of the differences between the brain and the CNNs, the Fisher-transformed correlation coefficient of each CNN was tested against those across the group of human participants using two-tailed *t* tests. Significant differences found are marked with asterisks for the subtraction-corrected (left set) and division-corrected (right set) brain data.

†*p* < 0.1, **p* < 0.05, ***p* < 0.01, and ****p* < 0.001.

correlation coefficient values). With the exception of image statistics whose correlation coefficient was marginally significantly greater than zero ($t_{(6)} = 2.38$, $p = 0.06$ for $d_{(s)}$ between, and $t_{(6)} = 2.42$, $p = 0.06$ for $d_{(d)}$ between, both two tailed and corrected for the four experiments included), all others were significantly greater than zero (t values > 2.86, p values < 0.025, two tailed and corrected). Confirming these results, the average of V1 to V4 was significantly lower than that of LOT and VOT (t values > 3.89, p values < 0.012 for both correction measures; all two tailed and corrected for the four experiments included). Together, these results show that not only was object information present in all the ventral regions examined here, but also, as visual processing progressed along the ventral visual pathway, object representation becomes more distinctive, with those in higher regions (i.e., VOT and LOT) being more distinctive than those in lower regions (i.e., V1 to V4). Moreover, similar results were found for both subtraction and division corrected Euclidean distance measures, indicating that the results did not depend on the specific correction procedure used.

To examine the presence and the strength of nonidentity information representation in the same brain regions, using the same split-half approach, we also calculated a corrected within-object Euclidean distance ($d_{(s)}$ within and $d_{(d)}$ within) in which we measured the distance between the same object category across the two values of a nonidentity feature (Fig. 1F, illustration). Following the same logic as we did before accessing identity representation, if a brain region contained distinctive representations for the different values of a nonidentity feature, then, when identity was held constant, a greater distance would be expected between two different values of a nonidentity feature than between the same value across the odd and even runs. This would result in $d_{(s)}$ within > 0 and $d_{(d)}$ within > 1. This was true for all four nonidentity features studied here in all the brain regions

examined (t values > 2.86, p values < 0.018, testing for values greater than zero or one, respectively, one tailed and corrected). Previously, we have only reported successful SVM decoding for SF (Vaziri-Pashkam et al., 2019), but did not examine the decoding of the other nonidentity features from these datasets. The present results show that all four types of nonidentity features were represented in all brain regions, consistent with other published studies (Hung et al., 2005; Eger et al., 2008; Konen and Kastner, 2008; Schwarzlose et al., 2008; Sayres and Grill-Spector, 2008; Kravitz et al., 2008, 2010; Carlson et al., 2011; Cichy et al., 2011, 2013; Hong et al., 2016; Reithler et al., 2017).

As with the identity features, we also tested whether the strength of nonidentity feature representations showed a linear increasing or decreasing trend across brain regions. Except for SF, which did not show a significant linear trend ($t_{(9)} = 0.58$, $p = 0.58$ for $d_{(s)}$ within, and $t_{(9)} = 0.52$, $p = 0.62$ for $d_{(d)}$ within, two tailed and corrected), all others showed a significantly negative trend along the cortical processing hierarchy (t values > 3.93, p values < 0.010, two tailed and corrected; Table 2, exact correlation coefficient values). Confirming these results, except for SF, the average of V1–V4 was significantly higher than that of LOT and VOT: t values > 5.83, p values < 0.002 for both correction measures, all two tailed and corrected; for SF: t values < 1.07, p values > 0.31 for both correction measures). Together, these results show that the representations of nonidentity features were present in all the ventral regions examined here, with the exception of SF, and that the representations of such information decreased significantly over the course of visual processing. As before, similar results were found for both subtraction- and division-corrected Euclidean distance measures, indicating that the results did not depend on the specific correction procedure used.

Overall, human ventral visual regions appear to amplify object identity and deemphasize nonidentity information over

Figure 4. Visualizing identity and position coding in human OTC and CNNs. Using MDS, the RDMs containing the different feature combinations were placed on 2D spaces with the distances among the different feature combinations approximating their similarities in the representational space.

the course of processing. For the 14 CNNs examined, we also calculated d_{between} and d_{within} for each sampled layer. Given the absence of measurement noise in CNNs, these distance measures were computed directly from the category responses from each layer without a split-half approach (Fig. 1E, illustration). For the 14 CNNs examined, object identity representation increased over the course of processing, consistent with the brain data (Figs. 2, 3). However, for the nonidentity features, CNNs overall showed a large increase followed by some decrease in their representations of the nonidentity features over the course of processing. To quantify these observations, we obtained the Pearson correlation coefficients between the Euclidean distance measures and the rank order of the CNN layers (Table 2, exact correlation coefficient values). To obtain a quantitative comparison between the brain and the CNNs, we tested whether or not the Fisher-transformed correlation coefficient value obtained from a CNN was significantly different from that of the brain across human participants. The results are reported in Table 2 for both subtraction- and division-corrected brain data (results are overall very similar for these two ways of correcting the brain data). Although the responses from a number of CNNs were indistinguishable from the brain for the representation of object identity and SF information, none showed a close match with the brain for the representation of position, size, and image statistics information.

Visualizing feature coding in human OTC and CNNs

While examining the Euclidean distance plots for an identity or a nonidentity visual feature alone tells us how the coding of that feature may change during the course of visual processing in the brain and CNNs, it does not inform us how they are coded with respect to each other and how they may jointly determine the representational structure over the course of visual processing. To directly visualize how object identity and nonidentity features may be represented together in a brain region, using the uncorrected Euclidean distances calculated from data averaged from all of the runs, we constructed a category RDM for each brain region (Kriegeskorte and Kievit, 2013; Fig. 1E). Using MDS (Shepard, 1980), we visualized this RDM by placing it on a 2D space with the distances among the different feature combinations approximating their similarities in representation (Figs. 4, 5, 6, 7). Here we used uncorrected distance measures from the average of all runs (Fig. 1E, illustration) rather than the corrected distances derived from the split-half approach (Fig. 1F, illustration) because of its simplicity and the fact that correction would largely serve as a scaling factor and would not drastically change the approximate depictions of the identity and nonidentity features within a given brain region in an MDS plot. A more precise and quantitative depiction of the representational space is provided by our identity dominance index in the next section.

Figure 5. Visualizing identity and size coding in human OTC and CNNs. Using MDS, the RDMS containing the different feature combinations were placed on 2D spaces with the distances among the different feature combinations approximating their similarities in the representational space.

A prominent feature of these MDS plots was the presence or absence of the separation between the object categories across the two values of each nonidentity feature. In V1, all four types of nonidentity features resulted in some separation in the representational space, with the separation being more striking because of position and size differences than because of image statistics and SF differences. Specifically, differences in position and size dominated the representational space and resulted in object categories segregated into two distinct and nonoverlapping groups, while differences in image statistics and SF, together with differences in object category, jointly shaped the representational space. As information processing ascended the visual pathway, in VOT, while the separation remained visible for the two Euclidean features position and size, object category representations became largely overlapping for the two non-Euclidean features image statistics and SF (Figs. 4, 5, 6, 7).

Applying the same procedure, we also visualized how object identity and nonidentity features may be represented together in a CNN layer. Just like in early visual areas, a separation for the different values of the four nonidentity features was present in the lower layers of all 14 CNNs (Figs. 4, 5, 6, 7). In the last fully connected layer and the pooling layer before that, however, not all CNNs behaved like the higher visual regions in their coding of position, size, and image statistics, and, critically, none resembled the brain in SF coding. The CNNs examined here thus do not appear to fully follow all the

feature coding characteristics of the human brain, especially in higher layers.

Quantifying the relative coding strength of identity and nonidentity features in human OTC

To quantify the relative coding strength of object identity and nonidentity features in a brain region and a CNN layer, using the subtraction corrected Euclidean distance measures $d_{(s)}^{\text{between}}$ and $d_{(s)}^{\text{within}}$, we constructed an identity dominance index (see Materials and Methods). With this index measure, an identity-only representation that disregards the nonidentity feature would have a $d_{(s)}^{\text{within}}$ value of 0 and an identity dominance value of 1. Conversely, a nonidentity-only representation that disregards identity would have a $d_{(s)}^{\text{between}}$ value of 0 and an identity dominance value of -1 .

As shown in Figure 8A, for all four types of nonidentity features examined, identity dominance varied significantly across brain regions (repeated-measures one-way ANOVA, F values > 6.79 , p values < 0.001), with a significantly positive linear trend across the visual processing hierarchy (t values > 2.91 , p values < 0.017 , two tailed and corrected for the four experiments included; Table 3, exact linear correlation values) and greater values for the higher than for the lower visual areas (the averages of LOT and VOT were all higher than those of V1 and V2, t values > 3.38 , p values < 0.008 , two tailed and corrected).

Figure 6. Visualizing identity and image statistics coding in human OTC and CNNs. Using MDS, the RDMs containing the different feature combinations were placed on 2D spaces with the distances among the different feature combinations approximating their similarities in the representational space.

Thus, object category coding strength increased over all four types of nonidentity features as information ascended the visual processing pathway. Nonetheless, differences existed among them, especially between the Euclidean and non-Euclidean features.

The two Euclidean features exhibited an overall similar response pattern. For position, identity dominance for early visual areas was very negative (the average of V1 and V2 was much lower than zero, $t_{(6)} = 146.86$, $p < 0.001$, two tailed and corrected), but increased to be close to 0 in higher visual areas (the average of VOT and LOT was still less than zero, $t_{(6)} = 8.32$, $p < 0.001$, two tailed and corrected; Fig. 8A). Identity dominance also differed among V1 to V4 (repeated-measures one-way ANOVA, $F_{(3,18)} = 7.90$, $p = 0.0014$) and linearly increased from V1 to V4 (with the averaged linear correlation coefficients being 0.66, and different from 0, $t_{(6)} = 4.03$, $p = 0.009$, two tailed and corrected). Pairwise tests between adjacent brain regions showed significant difference between V4 and LOT ($t_{(6)} = 3.82$, $p = 0.044$, two tailed and corrected) and marginally significant difference between V3 and V4 ($t_{(6)} = 3.14$, $p = 0.05$, two tailed and corrected). Other pairwise comparisons were not significant (t values < 2.18 , p values > 0.12 , two tailed and corrected). For size, identity dominance for early visual areas was also very negative (the average of V1 and V2 was much lower than zero, $t_{(6)} = 38.67$, $p < 0.001$, two tailed and corrected), but increased to be no

different from 0 in higher visual areas (the average of VOT and LOT to 0, $t_{(6)} = 0.61$, $p = 0.56$, two tailed and corrected; Fig. 8A). Like position, identity dominance also differed among V1 to V4 (repeated-measures one-way ANOVA, $F_{(3,18)} = 17.36$, $p < 0.001$) and linearly increased from V1 to V4 (with the averaged linear correlation coefficients being 0.80, and different from 0, $t_{(6)} = 9.23$, $p < 0.001$, two tailed and corrected). Pairwise tests between adjacent brain regions showed significant differences between V3 and V4, and V4 and LOT (t values > 3.44 , p values < 0.035 , two tailed and corrected). Differences between other adjacent brain regions were not significant (t values < 1.71 , p values > 0.23 , two tailed and corrected). These results indicated that position and size were much more prominent than objects in determining the representational space in early visual areas. The dominance of these two features over identity appeared to gradually decrease over the course of the visual processing. At higher visual regions LOT and VOT, position still dominated identity, whereas size and identity played a more or less similar role in shaping the representational space. Higher visual representations were thus never truly dominated by object identities, but rather maintained position and size information as a significant part of the object representation.

A different pattern emerged for the two non-Euclidean features. For image statistics, identity dominance started above zero in early visual areas (the average of V1 and V2 was above zero,

Figure 7. Visualizing identity and SF coding in human OTC and CNNs. Using MDS, the RDMs containing the different feature combinations were placed on 2D spaces with the distances among the different feature combinations approximating their similarities in the representational space.

$t_{(5)} = 5.15$, $p = 0.0048$; V1 alone was also above zero, $t_{(5)} = 2.60$, $p = 0.048$, two tailed and corrected) to become greatly above zero in higher visual areas (the average of LOT and VOT was much greater than zero, $t_{(5)} = 21.40$, $p < 0.001$, all two tailed and corrected; Fig. 8A). Identity dominance also differed among V1 to V4 (repeated-measures one-way ANOVA, $F_{(3,15)} = 15.85$, $p < 0.001$) and linearly increased from V1 to V4 (with the averaged linear correlation coefficients being 0.88, and different from 0, $t_{(5)} = 5.63$, $p = 0.005$, two tailed and corrected). Pairwise tests showed that the difference between V1 and V2, and that between V4 and LOT were both marginally significant ($t_{(5)} = 3.35$, $p = 0.051$, $t_{(5)} = 3.86$, $p = 0.060$, respectively, two tailed and corrected), with no difference between other adjacent regions (t values < 2.16 , p values > 0.13 , two tailed and corrected). Thus, object identity representation was more prominent than that of image statistics starting in early visual areas; over the course of visual processing, the strength of identity coding further increased and completely dominated representation at higher levels of visual processing. For SF, there was a similar overall trend of identity dominance going from 0 in early visual areas to significantly above zero in higher visual areas ($t_{(9)} = 0.62$, $p = 0.55$, and $t_{(9)} = 4.65$, $p = 0.0016$, respectively, for the difference between zero and the averages of V1 and V2 and that of LOT and VOT, two tailed and corrected). Identity dominance did not vary from V1 to V4 (repeated-measures one-way ANOVA,

$F_{(3,27)} = 1.73$, $p = 0.18$) and did not linearly increased from V1 to V4 (the averaged linear correlation coefficient was -0.069 and no different from 0, $t_{(9)} = 0.70$, $p = 0.50$, two tailed and corrected). Pairwise tests between adjacent brain regions revealed no significant differences (t values < 2.10 , p values > 0.16 , two tailed and corrected). Thus, unlike in image statistics, object identity and SF play a similar role in determining the representational space from V1 to V4. Given the presence of an overall difference and a positive linear trend when all six brain regions were examined together, and the presence of higher identity dominance for higher than for lower visual regions (see the summary statistics reported at the beginning of this section), it appears that identity dominates SF representation gradually over the course of visual processing, although none of the pairwise comparisons between adjacent brain regions reached significance. Overall, image statistics and SF appear to be either less or equally prominent as object identity in determining the representational space in early visual areas. As information ascends the visual processing hierarchy, object identity, rather than image statistics or SF, completely dominates the representational space.

To understand how the four types of nonidentity features may differ from each other, we also directly compared the identity dominance of these four features across the four experiments in lower visual regions (average of V1 and V2) and higher visual regions (average of LOT and VOT; Fig. 8C; data from the middle

Figure 8. Quantifying the relative coding strength of identity and nonidentity features in human OTC and CNNs using identity dominance indices. **A**, Identity dominance from human OTC comparing the coding of identity with each of the four nonidentity features from the 75 most reliable voxels across the odd and even runs and corrected for reliability using the subtraction method (Fig. 1*F*). To evaluate the consistency of the effect in the data, indices were also computed without correction for reliability (Fig. 1*E*) as correction could not be applied in the same way across the various splits of the data: from the 75 most reliable voxels and with distances measured across the odd and even runs, from odd runs only, from even runs only, and from all runs, and from all the voxels with all runs included in the initial Euclidean distance calculations. Very similar results were obtained in these different splits of the data, indicating the consistency of the effects in the data. Note that indices without reliability correction were compressed toward 0 because of the presence of measurement noise (which would generate positive distance measures even without positive feature representations). **B**, Identity dominance from the 14 CNNs comparing the coding of identity with each of the four nonidentity features, calculated following the procedures illustrated in Figure 1*E*. **C**, Direct comparisons of identity dominance indices across the four types of nonidentity features between the brain (reliability corrected using the subtraction method) and the 14 CNNs. For brain regions, lower, middle, and higher layers refer to the average of V1 and V2, the average of V3 and V4, and the average of LOT and VOT, respectively. For CNNs, lower, middle, and higher layers refer to the average of the first two layers, the average of all the middle layers excluding the first two and last two layers, and the average of the last two layers, respectively. The significance levels of the differences between the brain and each CNN are marked on top of each plot. $tp < 0.1$, $*p < 0.05$, $**p < 0.01$, and $***p < 0.001$.

region (the average of V3 and V4) were also included for completeness). Unpaired t tests revealed that, in lower visual regions, identity dominance was lower for position than size, lower for size than for both SF and image statistics (t values > 8.66 ,

p values < 0.001 , two tailed and corrected), and lower for SF than for image statistics (marginally significant, $t_{(14)} = 1.82$, $p = 0.091$, two tailed and corrected). In higher visual regions, identity dominance followed the same relative difference as

Table 3. The correlations between identity dominance and the rank order of brain regions or CNN layers in the four experiments

System name	Position	Size	Image statistics	SF
Brain (SE)	0.87 (0.02)	0.88 (0.02)	0.94 (0.02)	0.47 (0.15)
Alexnet	0.95**	0.95**	0.17***	−0.77***
Cornet-S	0.83	0.76**	0.64**	−0.43***
Densenet-201	0.87	0.81*	0.94	0.84*
Googlenet	0.90	0.88	0.90†	−0.65***
Inception_v3	0.89	0.95**	0.92	0.66
Inception-resnet_v2	0.83	0.73**	0.73**	0.62
Mobilenet_v2	0.72**	0.68***	0.70**	−0.49***
Resnet-18	0.81†	0.69***	0.89*	−0.35**
Resnet-50	0.81†	0.53***	0.80**	−0.64***
Resnet-101	0.80*	0.67***	0.83**	−0.25**
Squeezenet	0.84	0.76**	0.81**	−0.20**
Vgg-16	0.89	0.67***	0.80**	−0.38**
Vgg-19	0.86	0.54***	0.77**	−0.28**
Xception	0.84	0.64***	0.56***	0.19†
Resnet-50-IN	0.81†	0.53***	0.80**	−0.64***
Resnet-50-SIN	0.84	0.62***	0.88*	0.80†
Resnet-50-SININ	0.82†	0.65***	0.83*	0.23†
Resnet-50-SININ-IN	0.83†	0.69***	0.89*	−0.16**

A positive/negative correlation indicates that identity dominance increases/decreases over the course of processing. Differences in correlation between the brain and the CNNs are marked with asterisks. † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

in lower visual regions, with all the above pairwise comparisons reaching significance (t values > 4.80 , p values < 0.001 , two tailed and corrected).

To understand whether identity dominance measures were consistent across the odd and even halves of the runs, we calculated identity dominance measures separately for these datasets without reliability corrections following the procedures illustrated in Figure 1E (since our reliability correction required comparing data from the odd and even runs, and thus could not be directly applied to the results from just the odd or even runs). Although these uncorrected identity dominance measures underestimated the values compared with the corrected ones (see Materials and Methods for a more detailed explanation), the values obtained were similar when analyzed within just the odd runs, just the even runs, across the odd and even runs, and across all runs together (Fig. 8A). Similar results were also obtained when we include all runs and all voxels, rather than just the top 75 most reliable voxels, in our analysis (Fig. 8A). This indicates that our data were fairly consistent across the odd and even halves of the runs and across all the runs within the top 75 most reliable voxels and across all the runs when all the voxels were included.

Overall, we found that the representational strength of object identity information significantly increased over nonidentity information from lower to higher visual areas. Nevertheless, differences existed among the different nonidentity features, with identity dominating the two non-Euclidean features (image statistics and SF), but not the two Euclidean features (position and size) in higher OTC regions.

Quantifying the relative coding strength of identity and nonidentity features in CNNs

For the CNN layer output, as illustrated in Figure 1E, because of the absence of measurement noise, we used the uncorrected Euclidean distances and the same formula as before for the brain response to calculate identity dominance indices. Like the regions in OTC, for position, all CNNs tested showed very

negative identity dominance in the early layers (Fig. 8B). However, 12 of the 14 CNNs showed above zero identity dominance in the final layers and thus a dominance of identity over position coding not seen in the brain. For size, identity dominance of CNNs followed a similar trajectory as those of the brain, being very negative in the early layers and became close to 0 in the final layers. Nevertheless, 13 of the 14 CNNs showed a dip in the middle layers not seen in the brain. For image statistics, although, like the brain, identity dominance of a majority of the CNNs started close to 0 in early layers and then became greater than zero in the final layers, their identity dominance trajectories differed from the brain: instead of showing a gradual increase across the layers, CNNs showed either no increase across multiple layers, or a dip to below zero in the middle layers, before a rise was seen toward the end of the processing. Moreover, identity dominance was much lower in the CNNs than in the brain. For SF, identity dominance from CNNs were largely negative and none were above zero in the final layers like the brain. Overall, CNNs exhibited an overrepresentation of identity over position and an underrepresentation of identity over image statistics and SF from low to high layers. Even for size, where brain-like identity dominance was seen in the early and final layers, their trajectories, however, deviated from those of the brain.

To provide some quantifications of these observations, we assessed the presence of a linear trend for identity dominance from lower to higher layers for each CNN (Table 3, exact linear correlation values). We additionally tested whether the magnitude of the linear trend differed significantly from that of the brain (Table 3, significance levels). Among the 14 CNN tested, Googlenet and Inception_v3 were the top two CNNs which showed no difference from the brain for three of the four features.

Since a linear trend measurement did not directly compare the exact magnitude of identity dominance between the brain and CNNs, we additionally compared CNN identity dominance for the four types of nonidentity features directly with those obtained from the brain (Fig. 8C). With the exception of SF in the middle layers and size in the higher layers, differences in the absolute values of identity dominance were observed between the brain and CNNs for the other conditions in the lower, middle, and higher layers in a majority of the CNNs tested (Fig. 8C, statistics). Overall, except for size, differences between the brain and CNNs were greater for higher than for lower layers in CNNs, notably with an overrepresentation of identity over position and an underrepresentation of identity over image statistics and SF at the higher layers (Table 4, difference values and statistics).

The effect of training a CNN on original versus stylized ImageNet images

Although CNNs are believed to explicitly represent object shapes in the higher layers (Kriegeskorte, 2015; LeCun et al., 2015; Kubilius et al., 2016), emerging evidence suggests that CNNs may largely use local texture patches to achieve successful object classification (Ballester and de Araújo, 2016; Gatys et al., 2017) or local rather than global shape contours for object recognition (Baker et al., 2018). In a recent demonstration, CNNs were found to be poor at classifying objects defined by silhouettes and edges, and when texture and shape cues were in conflict, classifying objects according to texture rather than shape cues (Geirhos et al., 2019; but see Baker et al., 2018). However, when Resnet-50 was trained with stylized ImageNet images in which the original texture of every single image was replaced with the style of a

Table 4. Identity dominance differences (absolute values) between the CNNs and the brain for the lower layers/regions and higher layers/regions in the four experiments

System name	Position		Size		Image statistics		SF	
	Low/Hi	Diff	Low/Hi		Low/Hi	Diff	Low/Hi	Diff
Alexnet	0.43/0.39	−0.04	0.42/0.16	−0.26**	0.27/0.84	0.57***	0.35/0.79	0.44***
Cornet-S	0.47/0.75	0.28**	0.41/0.16	−0.25**	0.28/0.72	0.43***	0.28/0.68	0.40**
Densenet-201	0.40/0.84	0.44***	0.35/0.12	−0.23**	0.38/0.64	0.26**	0.33/0.48	0.16
Googlenet	0.47/0.75	0.28**	0.44/0.11	−0.33***	0.29/0.68	0.39***	0.27/0.62	0.35**
Inception_v3	0.23/0.86	0.63***	0.14/0.12	−0.03	0.43/0.57	0.14*	0.30/0.50	0.19†
Inception-resnet_v2	0.35/0.97	0.61***	0.37/0.11	−0.26***	0.26/0.47	0.22*	0.30/0.53	0.23*
Mobilenet_v2	0.46/0.65	0.19*	0.45/0.16	−0.30**	0.30/0.73	0.43***	0.23/0.65	0.42***
Resnet-18	0.51/0.80	0.29**	0.48/0.12	−0.35***	0.32/0.70	0.39***	0.26/0.61	0.35**
Resnet-50	0.44/0.60	0.17*	0.42/0.19	−0.23*	0.36/0.74	0.38**	0.27/0.71	0.45***
Resnet-101	0.41/0.83	0.42***	0.393/−0.12	−0.26**	0.39/0.70	0.27**	0.27/0.62	0.36**
SqueezeNet	0.44/0.32	−0.13	0.46/0.16	−0.30**	0.35/0.75	0.41***	0.28/0.56	0.28**
Vgg-16	0.57/0.73	0.17*	0.58/0.14	−0.45***	0.25/0.70	0.45***	0.26/0.56	0.30**
Vgg-19	0.57/0.60	0.03	0.57/0.15	−0.42***	0.26/0.70	0.43***	0.25/0.57	0.31**
Xception	0.28/0.93	0.65***	0.41/0.13	−0.28***	0.35/0.54	0.20*	0.26/0.51	0.25*
Resnet-50-IN	0.44/60	0.17*	0.42/0.19	−0.23*	0.36/0.74	0.38**	0.26/0.71	0.45***
Resnet-50-SIN	0.39/0.61	0.22*	0.37/0.18	−0.19*	0.42/0.47	0.06	0.27/0.46	0.19*
Resnet-50-SININ	0.39/0.84	0.45***	0.37/0.13	−0.24*	0.42/0.62	0.20*	0.26/0.52	0.26*
Resnet-50-SININ-IN	0.40/0.86	0.45***	0.38/0.12	−0.26**	0.41/0.64	0.23**	0.26/0.58	0.31**

To assess whether the absolute difference between the CNNs and brain changed from lower to higher layers/regions, difference values (Diff) were calculated by subtracting the lower layers/regions (Low) from the higher layers/regions (Hi) values. These difference values were tested against 0 at the participant-group level, and the results are marked with asterisks.

† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

randomly chosen painting, object classification performance significantly improved, relied more on shape than texture cues, and became more robust to noise and image distortions (Geirhos et al., 2019). It thus appears that a suitable dataset may overcome the texture bias in standard CNNs and allow them to use more shape cues.

Here we tested whether the relative coding strength of object identity and nonidentity features in a CNN may become more brain like when the CNN was trained with stylized ImageNet images. To do so, we examined the representations formed in Resnet-50 pretrained with the following three different procedures (Geirhos et al., 2019): trained only with the stylized ImageNet Images (RN50-SIN); trained with both the original and the stylized ImageNet Images (RN50-SININ); and trained with both sets of images and then fine-tuned with the stylized ImageNet images (RN50-SININ-IN). For comparison, we also included Resnet-50 trained with the original ImageNet images (RN50-IN) that we tested before. Across the different training procedures, while some measures improved, others stayed the same or became worse (Fig. 9, results, Table 2–Table 4, results). Overall, it was unclear that training with stylized images substantially improved performance for Resnet-50 across the board.

Discussion

Despite the usefulness of object identity and nonidentity features in vision and their joint coding in the primate visual brain, they have so far been studied relatively independently. Here we documented the coding strengths of object identity and nonidentity features together within a brain region and how they changed over the course of processing along the human ventral visual pathway. We examined the coding of object identity with four nonidentity features, including two Euclidean features (position and size) and two non-Euclidean features (basic image statistics and the SF content of an image). We additionally compared responses from the human brain with those from 14 CNNs pretrained for object recognition.

When features were examined separately, we found that identity representation increased, while nonidentity feature

representation decreased (with the exception of SF, which did not change), over the course of visual processing in the human brain. For the 14 CNNs examined, identity representations also increased over the course of processing, consistent with the brain data. However, nonidentity feature representations in CNN showed an initial large increase followed by a decrease at later stage of processing, different from brain responses.

Our results differed from those of Hong et al. (2016), who reported increased nonidentity feature representation in macaques and a CNN model over the course of processing. Hong et al. (2016) embedded objects in cluttered scenes, whereas we presented objects on uncluttered backgrounds. In cluttered scenes, objects may not be fully separated from the background without the formation of some object representation, resulting in a need to represent nonidentity features at higher levels of processing. Supporting this, Graumann et al. (2019) reported stronger position representation in higher ventral regions for objects in cluttered than in uncluttered scenes. Thus, scene clutter may account for the discrepancy in neural data between Hong et al. (2016) and our study. Hong et al. (2016) did find decreased nonidentity feature representation for a simple grating stimulus in both the brain and a CNN over the course of processing. However, this could be because of decreased representation for such a simple stimulus (i.e., with it being less preferred than complex visual objects at higher visual areas). Hong et al. (2016) did not test isolated complex objects in both systems. Our results showed that CNNs were unable to track the evolution of nonidentity feature representations in the human brain when complex objects were shown on uncluttered backgrounds.

In addition to examining object identity and nonidentity features separately, we also examined for the first time the relative coding strength of these features and its evolution along the ventral processing pathway. We found that identity coding significantly increased over all four types of nonidentity features from lower to higher visual areas, with identity dominating the non-Euclidean features (image statistics and SF) but not the Euclidean features (position and size) in higher OTC regions. Higher visual representations were thus never truly dominated

Figure 9. Results of Resnet-50 with different training regimes. Resnet-50 was pretrained either with the original ImageNet images (RN50-IN), the stylized ImageNet Images (RN50-SIN), both the original and the stylized ImageNet Images (RN50-SININ), or both sets of images, and then fine-tuned with the stylized ImageNet images (RN50-SININ-IN). **A**, The coding strengths of object identity and nonidentity features, with the between-object Euclidean distances and the within-object Euclidean distances measuring the coding strength of object identity and nonidentity features, respectively. Brain responses were corrected for reliability using the subtraction (Brain-s) method. **B**, Identity dominance measures. **C**, Direct comparisons of identity dominance across the four types of nonidentity features between the brain (reliability corrected using the subtraction method) and the CNNs. For brain regions, lower, middle, and higher layers refer to the average of V1 and V2, the average of V3 and V4, and the average of LOT and VOT, respectively. For CNNs, lower, middle, and higher layers refer to the average of the

by object identities, but rather by maintaining position and size information as part of the object representation. This is consistent with the existence of topographic maps in higher visual regions (Brewer et al., 2005) and the robust representation of position information in monkey IT cortex (DiCarlo and Maunsell, 2003; Hung et al., 2005; Hong et al., 2016) and in human VOT and LOT (Kravitz et al., 2008, 2010; Schwarzlose et al., 2008; Carlson et al., 2011; Cichy et al., 2011, 2013).

In our fMRI experiments, participants always attended identity by performing an identity one-back task. Does the relative coding strength of identity and nonidentity features depend on attention and task? In a previous study, when participants attended either object identity or color, we found highly similar and overlapping object category representational structures in OTC, but not in posterior parietal cortex (PPC; Vaziri-Pashkam and Xu, 2017, their experiment 3; color coding was not examined in the study). This suggests that the relative coding strength of identity and nonidentity features in PPC, but not those in OTC, may be modulated by attention and task. This would be consistent with the more adaptive nature of visual processing in PPC than in OTC (Xu, 2018a,b). Future research is needed to confirm this prediction.

Our measure of the relative coding strength of object identity and nonidentity features depended on the variation we introduced in each feature. For example, the relative coding strength of two similar object identities over two dissimilar object positions could be very different from that of two different object identities over two similar object positions. Because similar objects in one region may not be similar in another region, it would not have been possible to equate feature variations for all of the brain regions examined. Thus, we have chosen what we believed to be reasonable variations for each feature, including sampling of a wide range of real-world object categories, choosing position, size, and SF differences as far apart as possible given the constraint of our displays. The goal here was therefore not to measure the absolute feature coding bias in a brain region or a CNN layer, but, rather, to investigate for a set of reasonably chosen features, how the feature coding bias may change across visual areas and CNN layers. That being said, given the much greater position and size variance we experience in the real world than that manipulated here, the dominance of position and size over identity is likely underestimated here and is greater in all levels of visual processing during real-world visual perception.

first two layers, the average of all the middle layers excluding the first two and last two layers, and the average of the last two layers, respectively. The significance levels of the differences between the brain and each CNN are marked on top of the plots. $tp < 0.1$, $*p < 0.05$, $**p < 0.01$, and $***p < 0.001$.

Even within the context of the specific feature values chosen here, why would identity dominate the two non-Euclidean features, but not the two Euclidean features, examined at higher levels of ventral visual processing? One possibility is that Euclidean features are more essential in our direct interaction with the objects, such as in reaching and grasping, while the non-Euclidean features may be discarded once other object information is recovered. For example, once we identify an object and its position and size in a foggy viewing condition, information about how foggy it is may no longer be useful in guiding our further interaction with the object.

Compared with the brain, the 14 CNNs examined exhibited an overrepresentation of identity over position and an underrepresentation of identity over image statistics and SF from lower to higher layers. Even for sizes where brain-like responses were seen in the early and final layers, their trajectories, however, deviated from those of the brain. Overall, higher CNN layers exhibited a greater divergence from the corresponding brain regions than lower layers did in how features are represented with respect to each other. Similar results were obtained for a CNN trained with stylized object images that emphasized shape representation.

To increase SNR, we averaged responses from multiple exemplars from the same category and examined responses to object categories rather than individual exemplars. Previous research has reported similar category and exemplar response profiles in macaque IT cortex and human lateral occipital cortex with more robust responses generated by categories than individual exemplars because of increased SNR (Hung et al., 2005; Cichy et al., 2011). Rajalingham et al. (2018) found even better behavior–CNN correspondence at the category level, but not at the exemplar level. Comparing responses at the category level should have thus increased our chance of finding a close brain–CNN correspondence. Yet, we still found significant discrepancies between the brain and CNNs, more so at higher than lower levels of visual processing.

We included in this study both shallow and very deep CNNs. Deeper CNNs in general exhibit better object recognition performance (as evident from the ImageNet challenge results; Russakovsky et al., 2015) and can partially approximate the recurrent processing in ventral visual regions (Kar et al., 2019). The recurrent CNN we examined here, Cornet-S, explicitly models recurrent processing in ventral visual areas (Kar et al., 2019) and is considered by some as the current best model of the primate ventral visual regions (Kubilius et al., 2019). Yet, we observed similar performance between shallow and deep CNNs (e.g., Alexnet vs Googlenet), and the recurrent CNN did not perform better than the other CNNs.

The present results join those of other studies that reported differences between the CNNs and brain, such as the kind of features used in object recognition (Ballester and de Araújo, 2016; Ullman et al., 2016; Gatys et al., 2017; Baker et al., 2018; Geirhos et al., 2019), disagreement in representational structure between CNNs and brain/behavior (Kheradpisheh et al., 2016; Karimi-Rouzbahani et al., 2017; Rajalingham et al., 2018; Xu and Vaziri-Pashkam, 2021), the inability of CNN to explain >55% of the variance of macaque V4 and IT neurons (Cadieu et al., 2014; Yamins et al., 2014; Kar et al., 2019; Bao et al., 2020), and how the two systems handle adversarial images (Serre, 2019). Here we found limited brain–CNN correspondence in the coding strength of object identity and nonidentity features within a brain region/CNN layer and how they change over the course of visual processing. Thus, despite the success of CNNs in object

recognition, CNNs may not fully model the human visual system at their current states.

References

- Baker N, Lu H, Erlikhman G, Kellman PJ (2018) Deep convolutional networks do not classify based on global object shape. *PLoS Comput Biol* 14:e1006613.
- Ballester P, de Araújo RM (2016) On the performance of GoogLeNet and AlexNet applied to sketches. In: AAAI'16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp 1124–1128. Palo Alto, CA: Association for the Advancement of Artificial Intelligence.
- Bao P, She L, McGill M, Tsao DY (2020) A map of object space in primate inferotemporal cortex. *Nature* 583:103–108.
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* 57:289–300.
- Bettencourt KC, Xu Y (2016) Understanding location- and feature-based processing along the human intraparietal sulcus. *J Neurophysiol* 116:1488–1497.
- Brewer AA, Liu J, Wade AR, Wandell BA (2005) Visual field maps and stimulus selectivity in human ventral occipital cortex. *Nat Neurosci* 8:1102–1109.
- Cadieu CF, Hong H, Yamins DLK, 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.
- Carlson T, Hogendoorn H, Fonteijn H, Verstraten FAJ (2011) Spatial coding and invariance in object-selective cortex. *Cortex* 47:14–22.
- Cichy RM, Kaiser D (2019) Deep neural networks as scientific models. *Trends Cogn Sci* 23:305–317.
- Cichy RM, Chen Y, Haynes JD (2011) Encoding the identity and location of objects in human LOC. *Neuroimage* 54:2297–2307.
- Cichy RM, Sterzer P, Heinze J, Elliott LT, Ramirez F, Haynes JD (2013) Probing principles of large-scale object representation: category preference and location encoding. *Hum Brain Mapp* 34:1636–1651.
- 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. *Sci Rep* 6:27755.
- Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* 9:179–194.
- Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: IEEE Conference on CVPR Computer Vision and Pattern Recognition, 2009, CVPR 2009, pp 248–255. Piscataway, NJ: Institute of Electrical and Electronics Engineers.
- DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. *Trends Cogn Sci* 11:333–341.
- DiCarlo JJ, Maunsell JHR (2003) Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *J Neurophysiol* 89:3264–3278.
- DiCarlo JJ, Zoccolan D, Rust RC (2012) How does the brain solve visual object recognition? *Neuron* 73:415–434.
- Eger E, Kell CA, Kleinschmidt A (2008) Graded size sensitivity of object-exemplar-evoked activity patterns within human LOC subregions. *J Neurophysiol* 100:2038–2047.
- 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.
- Farah MJ (2004) Visual agnosia. Cambridge, MA: MIT.
- Gatys LA, Ecker AS, Bethge M (2017) Texture and art with deep neural networks. *Curr Opin Neurobiol* 46:178–186.
- Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Bren-del W (2019) ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv* 1811.12231.
- Goodale MA, Milner AD, Jakobson LS, Carey DP (1991) A neurological dissociation between perceiving objects and grasping them. *Nature* 349:154–156.
- Graumann M, Ciuffi C, Cichy RM (2019) Scene clutter and attention differentially affect object category and location representations [abstract]. *J Vis* 19:171a.

- Grill-Spector K, Kushnir T, Edelman S, Itzhak Y, Malach R (1998) Cue-invariant activation in object-related areas of the human occipital lobe. *Neuron* 21:191–202.
- Grill-Spector K, Kushnir T, Hendler T, Malach R (2000) The dynamics of object-selective activation correlate with recognition performance in humans. *Nat Neurosci* 3:837–843.
- Güçlü U, van Gerven MAJ (2017) Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *Neuroimage* 145:329–336.
- 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.
- Hong H, Yamins DLK, Majaj NJ, DiCarlo JJ (2016) Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat Neurosci* 19:613–622.
- Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005) Fast readout of object identity from macaque inferior temporal cortex. *Science* 310:863–866.
- Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. *Nat Neurosci* 8:679–685.
- Kar K, Kubilius J, Schmidt K, Issa EB, DiCarlo JJ (2019) Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat Neurosci* 22:974–983.
- Karimi-Rouzbahani H, Bagheri N, Ebrahimpour R (2017) Invariant object recognition is a personalized selection of invariant features in humans, not simply explained by hierarchical feed-forward vision models. *Sci Rep* 7:14402.
- Khaligh-Razavi S-M, Kriegeskorte N (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput Biol* 10:e1003915.
- Kheradpisheh SR, Ghodrati M, Ganjtabesh M, Masquelier T (2016) Deep networks can resemble human feed-forward vision in invariant object recognition. *Sci Rep* 6:32672.
- Konen CS, Kastner S (2008) Two hierarchically organized neural systems for object information in human visual cortex. *Nat Neurosci* 11:224–231.
- Kourtzi Z, Kanwisher N (2000) Cortical regions involved in perceiving object shape. *J Neurosci* 20:3310–3318.
- Kravitz DJ, Vinson LD, Baker CI (2008) How position dependent is visual object recognition? *Trends Cogn Sci* 12:114–122.
- Kravitz DJ, Kriegeskorte N, Baker CI (2010) High-level visual object representations are constrained by position. *Cereb Cortex* 20:2916–2925.
- 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, Kievit RA (2013) Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn Sci* 17:401–412.
- Kubilius J, Bracci S, Op de Beeck HP (2016) Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput Biol* 12:e1004896.
- Kubilius J, Schrimpf M, Hong H (2019) Brain-like object recognition with high-performing shallow recurrent ANNs. In: *NeurIPS | 2019, Thirty-Third Conference on Neural Information Processing Systems*. San Diego: Neural Information Processing Systems.
- LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444.
- Malach R, Reppas JB, Benson RR, Kwong KK, Jiang H, Kennedy WA, Ledden PJ, Brady TJ, Rosen BR, Tootell RB (1995) Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proc Natl Acad Sci U S A* 92:8135–8139.
- O'Connell TP, Chun MM (2018) Predicting eye movement patterns from fMRI responses to natural scenes. *Nat Commun* 9:5159.
- Orban GA, Van Essen D, Vanduffel W (2004) Comparative mapping of higher visual areas in monkeys and humans. *Trends Cogn Sci* 8:315–324.
- Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ (2018) Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J Neurosci* 38:7255–7269.
- Reithler J, Peters JC, Goebel R (2017) Characterizing object- and position-dependent response profiles to uni- and bilateral stimulus configurations in human higher visual cortex: a 7T fMRI study. *Neuroimage* 152:551–562.
- Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet large scale visual recognition challenge. *Int J Comput Vis* 115:211–252.
- Sayres R, Grill-Spector K (2008) Relating retinotopic and object-selective responses in human lateral occipital cortex. *J Neurophysiol* 100:249–267.
- Schwarzlose RF, Swisher JD, Dang S, Kanwisher N (2008) The distribution of category and location information across object-selective regions in human visual cortex. *Proc Natl Acad Sci U S A* 105:4447–4452.
- Sereno MI, Dale AM, Reppas JB, Kwong KK, Belliveau JW, Brady TJ, Rosen BR, Tootell RB (1995) Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science* 268:889–893.
- Serre T (2019) Deep learning: the good, the bad, and the ugly. *Annu Rev Vis Sci* 5:399–426.
- Shepard RN (1980) Multidimensional scaling, tree-fitting, and clustering. *Science* 210:390–398.
- Swisher JD, Halko MA, Merabet LB, McMains SA, Somers DC (2007) Visual topography of human intraparietal sulcus. *J Neurosci* 27:5326–5337.
- Tarhan L, Konkle T (2020) Reliability-based voxel selection. *Neuroimage* 207:116350.
- Ullman S, Assif L, Fetaya E, Harari D (2016) Atoms of recognition in human and computer vision. *Proc Natl Acad Sci U S A* 113:2744–2749.
- Vaziri-Pashkam M, Xu Y (2017) Goal-directed visual processing differentially impacts human ventral and dorsal visual representations. *J Neurosci* 37:8767–8782.
- Vaziri-Pashkam M, Xu Y (2019) An information-driven 2-pathway characterization of occipito-temporal and posterior parietal visual object representations. *Cereb Cortex* 29:2034–2050.
- Vaziri-Pashkam M, Taylor J, Xu Y (2019) Spatial frequency tolerant visual object representations in the human ventral and dorsal visual processing pathways. *J Cogn Neurosci* 31:49–63.
- Willenbockel V, Sadr J, Fiset D, Horne GO, Gosselin F, Tanaka JW (2010) Controlling low-level image properties: the SHINE toolbox. *Behav Res Methods* 42:671–684.
- Williams MA, Dang S, Kanwisher NG (2007) Only some spatial patterns of fMRI response are read out in task performance. *Nat Neurosci* 10:685–686.
- Xu Y (2018a) A tale of two visual systems: invariant and adaptive visual information representations in the primate brain. *Annu Rev Vis Sci* 4:311–336.
- Xu Y (2018b) The posterior parietal cortex in adaptive visual processing. *Trends Neurosci* 41:806–822.
- Xu Y, Vaziri-Pashkam M (2021) Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nat Commun* 12:2065.
- Yamins DLK, DiCarlo JJ (2016) Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci* 19:356–365.
- Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci U S A* 111:8619–8624.