PNAS RESEARCH ARTICLE | NEUROSCIENCE

Texture-like representation of objects in human visual cortex

Akshay V. Jagadeesh[a,b,1] and Justin L. Gardner[a,b]

Edited by J. Anthony Movshon, New York University, New York, NY; received August 30, 2021; accepted March 15, 2022

The human visual ability to recognize objects and scenes is widely thought to rely on representations in category-selective regions of the visual cortex. These representations could support object vision by specifically representing objects, or, more simply, by representing complex visual features regardless of the particular spatial arrangement needed to constitute real-world objects, that is, by representing visual textures. To discriminate between these hypotheses, we leveraged an image synthesis approach that, unlike previous methods, provides independent control over the complexity and spatial arrangement of visual features. We found that human observers could easily detect a natural object among synthetic images with similar complex features that were spatially scrambled. However, observer models built from BOLD responses from category-selective regions, as well as a model of macaque inferotemporal cortex and Imagenet-trained deep convolutional neural networks, were all unable to identify the real object. This inability was not due to a lack of signal to noise, as all observer models could predict human performance in image categorization tasks. How then might these texture-like representations in category-selective regions support object perception? An image-specific readout from category-selective cortex yielded a representation that was more selective for natural feature arrangement, showing that the information necessary for natural object discrimination is available. Thus, our results suggest that the role of the human category-selective visual cortex is not to explicitly encode objects but rather to provide a basis set of texture-like features that can be infinitely reconfigured to flexibly learn and identify new object categories.

ventral visual stream | deep neural networks | object perception | texture representation | BOLD

Significance

Humans are exquisitely sensitive to the spatial arrangement of visual features in objects and scenes, but not in visual textures. Category-selective regions in the visual cortex are widely believed to underlie object perception, suggesting such regions should distinguish natural images of objects from synthesized images containing similar visual features in scrambled arrangements. Contrarily, we demonstrate that representations in category-selective cortex do not discriminate natural images from feature-matched scrambles but can discriminate images of different categories, suggesting a texture-like encoding. We find similar insensitivity to feature arrangement in Imagenet-trained deep convolutional neural networks. This suggests the need to reconceptualize the role of category-selective cortex as representing a basis set of complex texture-like features, useful for a myriad of behaviors.

Images engineered to have the same complex visual features as natural images can appear metameric (1, 2), that is, perceptually indistinguishable although physically different from the original natural image. In particular, by matching image statistics (3, 4), including the pairwise correlations of orientation and spatial frequency filters (5) or the pairwise inner products of feature maps from Imagenet-trained deep convolutional neural networks (dCNNs) (6–8), it is possible to synthesize images which appear indistinguishable from the corresponding natural image, despite having a spatially scrambled arrangement of features (9, 10). This metamer synthesis approach is particularly effective for visual textures, such as bark, gravel, or moss, which contain complex visual features that are largely homogeneous over space (11–15). The synthesis approach has been used to study the phenomenon of crowding (16) in peripheral vision (10, 17, 18), where observers can fail to bind visual features to corresponding objects when presented in visual clutter (19, 20). The neural representation of complex visual features has also been studied with texture synthesis approaches (21–25).

However, images which contain inhomogeneous visual features, such as those of objects or natural scenes, are perceptually distinct from synthesized images which contain the same complex visual features but in scrambled spatial arrangements (14, 26, 27). This suggests that the underlying neural representation of objects and scenes, in contrast to that of textures, is sensitive to the particular spatial arrangement of features found in objects and scenes in the natural world.

The lateral occipital and ventral visual cortex of humans are potential cortical substrates for such representations which distinguish objects and natural scenes from synthesized, scrambled counterparts. Studies of early visual cortical representations suggest that sensitivity to the midlevel visual features contained in texture images can be found in areas V2 (21–23) and V4 (24, 25, 28), while category-selective representations in higher-level visual cortical areas within lateral occipital cortex (LO) and ventral temporal cortex (VTC) are informative for decoding object categories (29–35) and predicting object categorization behavior (30, 36, 37). Thus, one might hypothesize that cortical representations in category-selective regions underlie the perceptual ability to discriminate natural object images from synthesized images containing spatially scrambled arrangements of complex visual features (38–43).

Author affiliations: [a]Department of Psychology, Stanford University, Stanford, CA 94305; and [b]Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA 94305

Author contributions: A.V.J. and J.L.G. designed research; A.V.J. performed research; A.V.J. and J.L.G. contributed new reagents/analytic tools; A.V.J. analyzed data; and A.V.J. and J.L.G. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: akshayj@stanford.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2115302119/-/DCSupplemental.

Published April 19, 2022.

However, evidence of texture bias in dCNN models (44–46) of the ventral visual stream suggests that it is possible to encode features useful for discriminating visual object categories without explicitly encoding the global arrangement of those features. Imagenet-trained (47) dCNNs have achieved state-of-the-art performance at modeling ventral visual cortical representations (48–52), but, unlike human perception, these dCNN models are biased to classify image category based on texture rather than shape information (53–56). That is, when texture and shape information are put into conflict, Imagenet-trained dCNNs frequently report the image category based on texture information, whereas humans reliably report the category consistent with the shape. The evidence of texture bias in dCNNs but not in human perception suggests two possible hypotheses: Either dCNNs may not be accurate models of the primate ventral visual cortex, particularly with regard to the processing of shape information, or the representations in the ventral visual cortex are also texture-like (57) and are therefore insufficient to account for humans' perceptual ability to discriminate natural scenes from synthesized images containing the same complex visual features in scrambled arrangement.

One approach to address whether the human ventral visual cortex explicitly represents the spatial arrangement of features that defines an object or natural scene is to examine the representational dissimilarity between natural images and synthesized images that have the same visual features but are scrambled. Prior studies employing this scrambling technique have been instrumental for the discovery of complex feature selectivity and invariance in high-level visual cortical areas (28, 57–61). However, these studies have often used methodologies for scrambling visual features which confound sensitivity to complex visual features with sensitivity to the spatial arrangement of those features. For example, Grill-Spector et al. (58) reported object sensitivity in LO by contrasting the blood oxygen level–dependent (BOLD) response to natural object images with the response to grid-scrambled images, a scrambling approach which breaks up complex visual features. Similarly, Rust and DiCarlo (28) demonstrated enhanced selectivity and tolerance to objects in macaque inferotemporal (IT) cortex by comparing neural responses to natural object images with responses to synthesized scrambles containing only low- and midlevel visual features. Thus, it remains undetermined whether cortical representations in category-selective regions of the visual cortex are merely sensitive to the presence of complex visual features or whether they are also sensitive to the spatial arrangement of those features. Furthermore, prior research has primarily examined the magnitude of response averaged across entire cortical areas (21, 58–60), rather than the multivariate patterns of population activity that might provide much richer featural representations. Finally, this work has largely overlooked the link between these cortical representations and the perceptual behaviors which they support.

In the present study, we compared the ability of human observers, dCNN models, and cortical representations in category-selective regions of the human visual cortex to discriminate natural images from synthesized images containing scrambled complex features. We sought to avoid the confound between feature complexity and spatial arrangement by adapting a hierarchical, spatially constrained image synthesis algorithm that allows control over the complexity of features and the spatial extent over which those features can be scrambled in the synthesized output (6, 7). In an oddity detection task, we found that human observers readily discriminated natural from synthetic images of objects, although performance was sensitive to the complexity and spatial arrangement of visual features. In stark contrast, Imagenet-trained dCNN models performed at chance level when the synthetic images contained complex visual features, regardless of the spatial arrangement of those features. Human visual cortical responses measured with BOLD imaging, as well as models of macaque IT neurons, similarly performed poorly in detecting the natural image. These results were not simply due to measurement noise, as we found that visual cortical BOLD responses could match human performance in category discrimination and that patterns of BOLD activity for different synthetic and natural images were reliable and discriminable. Rather, the representational distance between natural and synthetic images was not significantly greater than the distance between synthetic images, suggesting a texture-like representation of objects which does not preferentially represent the natural spatial arrangements of object features.

Results

Human Perceptual Sensitivity for Natural vs. Synthesized Images. To assess perceptual sensitivity to the complexity and spatial arrangement of features, we used an oddity detection task (see Fig. 2*A*), in which subjects were presented with three images on each trial (one natural, two synthesized) and asked to choose the odd one out, that is, the image which appeared the most different from the others. Images subtended 8° in diameter and were centered 6° from the center of fixation. On each trial, both synthesized images ("synths") were matched to the features of the natural image at a particular level of feature complexity and spatial arrangement constraint. To synthesize images, we extracted feature maps from various VGG-19 layers (e.g., pool1, pool2, pool4) in response to a natural image, then computed the pairwise dot products between each pair of feature maps (Gramian) within subregions of different levels of spatial constraint (from least spatially constrained, 1×1, where features could be spatially scrambled across the whole image to most spatially constrained, 4×4, in which features could only be scrambled within 16 subregions of the image) (Fig. 1*A*). Then, we iteratively optimized the pixels of a randomly initialized white noise image to minimize the mean squared error between its Gramian feature representation and that of the original. We will refer to the feature complexity of a synth to indicate the latest dCNN layer whose features were extracted and matched to the original. We will use the term spatial constraint or spatial arrangement to indicate the size of the spatial pooling regions within which those features were matched. Across trials, we varied the feature complexity of the synths as well as the degree to which the spatial arrangement of those features was constrained (Fig. 1*B*). We will use the term image class to refer to the set of images including a given natural image and all its feature-matched synths.

In the oddity detection task, we found that human observers were less able to detect the natural image among synths with more-complex features. That is, we examined detection performance as a function of the feature complexity of the synths, pooled across all observers and averaged across all spatial constraints. We reasoned that this would be informative of which features are utilized in the perception of natural images of objects. We found that increasing the feature complexity of the synths resulted in a significant decline (linear mixed effects model: $b = -0.103$, SE = 0.007, $P < 0.001$, 95% CI = $[-0.116, -0.090]$, $n = 87$) in the proportion of trials where the natural image was chosen as the oddity (Fig. 2*C*; note the downward slope of the purple line), suggesting that human

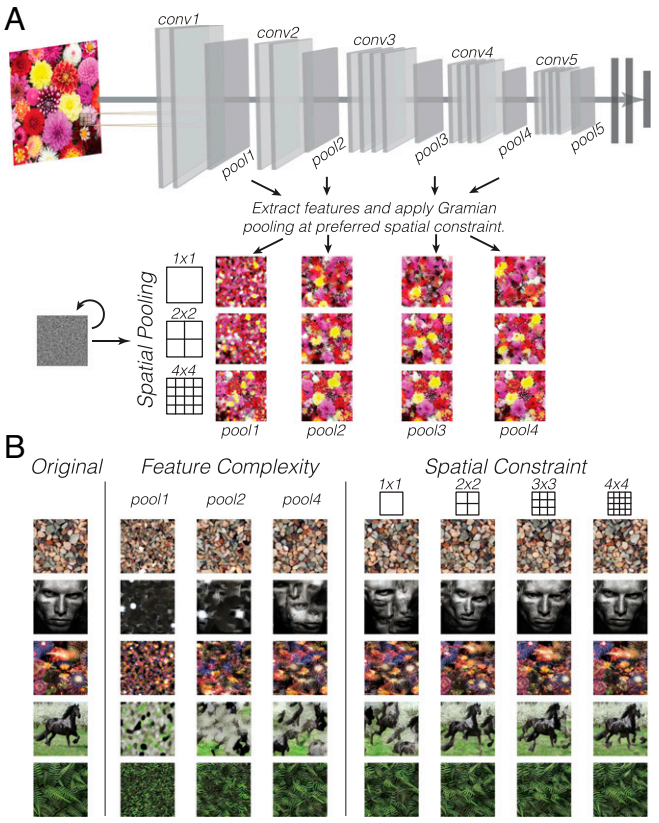

Fig. 1. Image synthesis algorithm and example synths. (*A*) Schematic of deep image synthesis algorithm. We pass a natural image into an Imagenet-trained VGG19 model and extract intermediate layer activations from layers pool1, pool2, and pool4. Then, we compute the correlation between pairs of feature maps within a layer (Gramian) constrained within spatial pooling regions (1 × 1, 2 × 2, 3 × 3, or 4 × 4). To synthesize, we iteratively update the pixels of a random seed image using gradient descent to match the spatially pooled Gramians from the original image. (*B*) Example natural images and feature-matched synths, varying in feature complexity (columns 2 to 4, fixed at 1 × 1 spatial constraint), and varying in spatial constraint (columns 5 to 8, fixed at pool4 complexity).

observers are perceptually sensitive to the complexity of visual features in object images.

We found that human observers were also less able to detect the natural image among synths with more constrained feature arrangement. To arrive at this conclusion, we assessed perceptual sensitivity for the spatial arrangement of visual features by analyzing oddity detection performance as a function of the spatial constraints in the synths, pooled across all observers, fixing feature complexity at the highest level (pool4). We reasoned that, if observers were no less able to detect the natural image among synths whose features were constrained within small subregions of the image (e.g., 4 × 4 condition) as compared to synths whose features were scrambled across the entirety of the image (1 × 1 condition), this would demonstrate that humans are insensitive to the arrangement of complex visual features. We found, instead, that the proportion of trials where subjects selected the natural image as the odd one out significantly decreased (linear mixed effects model: $b = -0.087$, SE = 0.011, $P < 0.001$, 95% CI = [−0.108, −0.066]) as the arrangement of object features was more strongly constrained (Fig. 2*D*; note the downward slope of the purple line). This pattern of behavior was consistent regardless of whether behavioral data were collected in-laboratory with fixation enforced (*SI Appendix*, Fig. S3) or online. These findings demonstrate that human observers' perception of objects is not only sensitive to the presence of the complex visual features that make up an object but also to the particular spatial arrangement of those features.

dCNN Observer Models. To compare the behavior of dCNN models to that of human observers, we constructed an observer model that uses dCNN features to perform the oddity detection task (Fig. 2*B*). On each trial, we first extracted a feature vector corresponding to each image, from the last convolutional layer of an Imagenet-trained dCNN. We then computed the Pearson distance between each pair of feature vectors. To determine which image was most different from the other two, we computed the dissimilarity of each image by averaging the distance from each image's feature vector to each of the two other images' feature vectors. These dissimilarities were then transformed into choice probabilities using a softmax function, with one free parameter controlling how sensitive the model is to dissimilarity differences, which was estimated to maximize the likelihood of human observers' choices. We evaluated the performance of five different Imagenet-trained dCNNs: VGG-19 (62) (the same model used for image synthesis), CORnet-Z (63), VGG-16 (62), ResNet-18 (64), and AlexNet (46). We hypothesized that, if dCNNs are an accurate model of human object perception, they ought to match human performance at this oddity detection task, in terms of sensitivity to both feature complexity and spatial arrangement.

Analysis of the dCNN observer models' performance showed that, when synths contained more-complex features, the models were less able to identify the natural image as the odd one out. That is, we compared each model's performance to human oddity detection performance as a function of the feature complexity of the synthesized images contained, averaged across all spatial constraints. Like human observers, almost all models were very likely to select the natural image when presented among synths containing only low-level visual features (Fig. 2*C*, pool1). However, increasing the feature complexity of the synths resulted in a steep decline in the probability of selecting the natural image (linear mixed effects model: $b = -0.249$, SE = 0.001, $P < 0.001$, 95% CI = [−0.251, −0.247]), such that all dCNN observer models were not significantly more likely than chance ($b = -0.038$, SE = 0.012, $P = 0.998$, 95% CI = [−0.061, −0.014]) to identify the natural object image among two synths containing complex visual features (Fig. 2*C*, pool4). This contrasts with human observers, whose frequency of selecting the natural image as the oddity did decline with increasing feature complexity but was still significantly above chance even at the highest level of feature complexity (Fig. 2*C*; purple line). Taken together, these observations suggest that dCNN observer models are sensitive to the presence of complex visual features in objects, but not to their spatial arrangement.

Indeed, we found that dCNN observer models failed to detect the natural image among synths containing complex visual features, regardless of how spatially scrambled those features were. We examined oddity detection performance as a function of spatial pooling region size, for the complex (pool4) feature condition alone, to isolate the effect of spatial arrangement when fixing feature complexity at the highest possible level (Fig. 2*D*). Whereas human observers detected the natural image most frequently when feature arrangement was least constrained (Fig. 2*D*, 1 × 1; note the purple line) and less frequently as feature arrangement became more constrained, all five dCNN observer models were not significantly more likely than chance to select the natural image in any condition

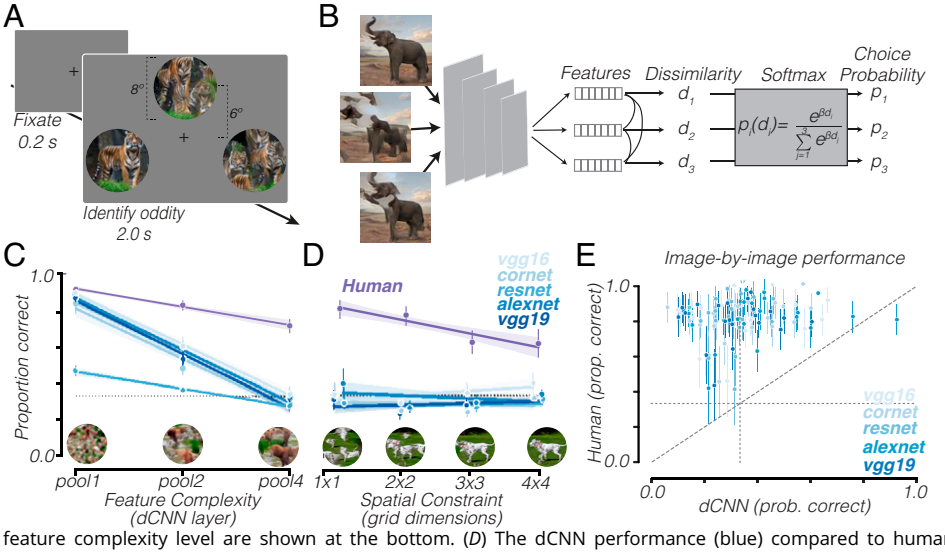

Fig. 2. Human perception of objects is sensitive to feature complexity and spatial arrangement while dCNN observer models are insensitive to spatial arrangement. (*A*) Oddity detection task. Subjects saw three images, one natural and two synths, and chose the odd one out by key press. (*B*) Schematic of dCNN observer model fit to oddity detection task. Features are activations for each image extracted from the last convolutional layer of a dCNN. The Pearson distances between each image's feature vector is computed (d_1 to d_3) and converted into choice probabilities (p_1 to p_3) using a softmax function with free parameter β which controls how sensitive the model is to feature dissimilarity. (*C*) The dCNN performance (blue) compared to human performance (purple) as a function of synths' feature complexity, averaged across all observers, images, and spatial constraint levels. Performance indicates the proportion of trials in which the natural image was chosen as the oddity. Example synths from each feature complexity level are shown at the bottom. (*D*) The dCNN performance (blue) compared to human performance (purple) as a function of synths' spatial constraints, across all observers and images for pool4 feature complexity. Example synths from each spatial constraint level are shown at the bottom. (*E*) The dCNN performance vs. human performance, image by image, for 1 × 1 pool4 condition. Vertical and horizontal dotted lines in *C*–*E* represent chance level. Diagonal dashed line in *E* is line of equality. Error bars in *C*–*E* indicate bootstrapped 95% CIs across trials. prob., probability; prop., proportion.

($b = -0.038$, SE = 0.012, $P = 0.998$, 95% CI = [−0.061, −0.014]) and were insensitive to variation in spatial arrangement ($b = 0.003$, SE = 0.001, $P = 0.02$, 95% CI = [0.000, 0.005]) (Fig. 2*D*; note the flat blue lines). Thus, unlike human observers, who reliably report that the natural image stands out among synthesized images, dCNN models are unable to identify the natural image as the odd one out even when presented among images whose features are scrambled across the entire image (1 × 1). Across a large sample of object image classes (Fig. 2*E*), we found that humans were significantly more likely than all Imagenet-trained dCNN models we tested to choose the natural image as the oddity when presented among synths containing scrambled (1 × 1) complex (pool4) visual features (VGG16: $t = 12.84$, $P < 0.001$; VGG19: $t = 14.93$, $P < 0.001$; CORnet: $t = 14.04$, $P < 0.001$; ResNet18: $t = 17.04$, $P < 0.001$; AlexNet: $t = 9.47$, $P < 0.001$). Thus, given that dCNNs were at floor performance regardless of how well constrained the arrangement of features was, we hypothesized that what distinguishes humans' perception of objects from that of dCNN observer models is not the ability to detect complex visual features but rather a selectivity for the particular spatial arrangement of features found in natural objects.

Category Oddity Detection Task. These behavioral and modeling results demonstrate that dCNN features are insensitive to the arrangement of complex features, suggesting that dCNNs do not represent objects but instead contain a texture-like representation of disjointed complex visual features. One implication of this is that object categorization, the task that Imagenet-trained dCNNs are optimized to perform, does not require an explicit representation of objects, merely a representation of the complex features that make up objects. To test this, we compared the performance of human observers and dCNN observer models in a category oddity detection task (Fig. 3*A*). On each trial, observers saw three different natural images—two of which belonged to the same category, while the third image contained an image from a different category—and were instructed to choose the odd one out. To make this task comparable to the previous task, subjects were instructed to select the image which appeared most different from the other two and were not explicitly directed to choose the image which belonged to a different category. We hypothesized that human observers' behavior on this category oddity detection task would be predicted well by the performance of dCNN observer models.

Indeed, we found that human observers' behavioral performance at category oddity detection was matched by the performance of dCNN observer models. We compared the frequency with which dCNNs selected the image belonging to the odd category out to that of human observers. We found that, across a variety of object categories, there was no significant difference ($t = 1.16$, $P = 0.26$, $n = 16$) in performance between human observers and dCNN observers (Fig. 3*B*). This suggests that dCNN representations, although insensitive to feature arrangement, are informative for discriminating categories that differ in their visual features.

Human Visual Cortex. To assess the ability of the human visual cortex to discriminate between natural and synthesized object images, we measured BOLD responses in the brains of seven human observers, while subjects passively viewed natural and synthesized (1 × 1, pool4) images from 10 different image classes (65). Images were presented for 4 s, subtending 12°, centered 7° to the left and right of fixation (Fig. 4*A*). We estimated trial-averaged responses to each individual image using a generalized linear model (66, 67). We analyzed data from 13 visual cortical areas, including V1, V2, V3, and hV4, which were retinotopically defined (68), midfusiform (mFus), posterior fusiform (pFus), inferior occipital gyrus (IOG), transverse occipital sulcus (TOS), and collateral sulcus (CoS), which were functionally defined using a functional localizer (69), and lateral occipital cortex (LO), ventral visual cortex (VVC), posterior IT cortex (PIT), and ventromedial visual area (VMV), which were anatomically defined (70). We use the term early visual cortex to refer to V1, V2, V3, and hV4, and use the terms category-selective regions or category-selective cortex to refer to mFus, pFus, IOG, TOS, CoS, VVC, PIT, LO, and VMV.

We were able to measure reliable patterns of BOLD activity in these 13 cortical areas in every subject. We estimated the split-half reliability of each voxel as the correlation between its responses, estimated on two different halves of image

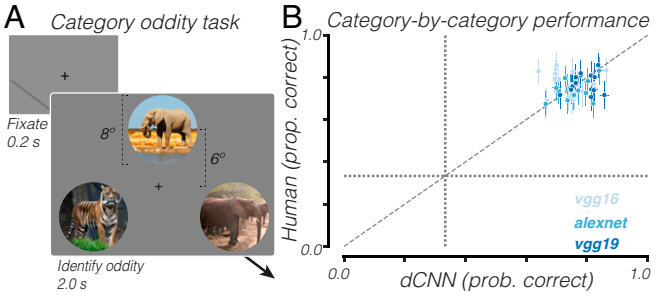

Fig. 3. The dCNN observer models match human performance in a category oddity task. (*A*) Category oddity detection task. On each trial, subjects saw three natural images—two of the same category, one from a different category—and chose the odd one out by key press. (*B*) The dCNN vs. human performance, category by category. Proportion correct denotes the proportion of trials in which the odd category out was correctly chosen. Diagonal dashed line is line of equality. Vertical/horizontal dotted lines represent chance level. Error bars indicate bootstrapped 95% CIs across trials. prob., probability; prop., proportion.

presentations. In each visual area, we selected the 100 voxels with the highest split-half reliability, to ensure our findings reflected a reliable signal rather than just noise. The mean reliability across all subjects of the 100 selected voxels in each of these visual areas exceeded the chance level derived from permutation testing in all analyzed visual areas ($R_{V1} = 0.70$; $R_{V2} = 0.68$; $R_{V3} = 0.63$; $R_{hV4} = 0.56$; $R_{mFus} = 0.30$; $R_{pFus} = 0.47$; $R_{IOG} = 0.42$; $R_{CoS} = 0.45$; $R_{TOS} = 0.44$; $R_{LO} = 0.58$; $R_{VVC} = 0.33$; $R_{PIT} = 0.33$; $R_{VMV} = 0.44$).

To determine whether visual cortical responses could support behavior in the oddity detection task, we constructed an observer model that used BOLD responses to choose the image which was most different (Fig. 4*B*). On each trial, we first extracted a vector of voxel responses from a given visual area to each of the three images that the human saw, then computed

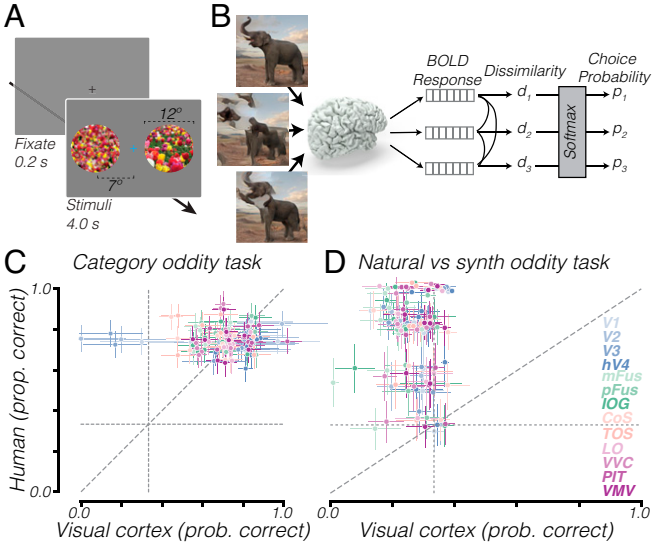

Fig. 4. BOLD responses match human performance at category discrimination but not natural-vs.-synth discrimination. (*A*) Stimulus design for BOLD imaging experiment. Natural and synthesized images were presented to the left and right of fixation while subjects performed a color discrimination task at the fixation cross. (*B*) Schematic of cortical observer model which uses BOLD responses to perform oddity detection task. Conventions are similar to Fig. 1*B*. (*C*) Cortical observer model vs. human performance on category oddity detection task, category by category. (*D*) Cortical observer models vs. human performance, image by image, on natural–synth discrimination task. Diagonal dashed line is line of equality. Vertical/horizontal dotted lines represent chance level. Error bars indicate bootstrapped 95% CIs across trials. prob., probability; prop., proportion.

the Pearson distance between each response vector, averaged together each pair of distances to estimate the representational dissimilarity of each image (mean distance from other two images), and then transformed the dissimilarities into choice probabilities using a softmax function. We evaluated the ability of this cortical observer model to identify the odd image out in both the category oddity detection task, where two objects of the same category were presented alongside a third object of a different category, and the natural-vs.-synth oddity detection task, where two synths containing scrambled complex visual features were presented alongside a natural image.

In a category oddity detection task, human observers' behavioral performance was matched by that of a cortical observer model using BOLD responses from the human visual cortex. We found that there was no significant difference (V1: $t = -1.49$, $P = 0.17$; V2: $t = -0.75$, $P = 0.47$; V3: $t = -1.04$, $P = 0.32$; hV4: $t = 4.47$, $P = 0.02$; mFus: $t = -4.05$, $P = 0.03$; pFus: $t = -0.88$, $P = 0.40$; IOG: $t = -1.18$, $P = 0.27$; CoS: $t = -4.31$, $P = 0.02$; TOS: $t = -1.46$, $P = 0.18$; LO: $t = -1.94$, $P = 0.09$; VVC: $t = -1.17$, $P = 0.29$; PIT: $t = 0.74$, $P = 0.48$; VMV: $t = -1.63$, $P = 0.14$) in the likelihood of selecting the odd category out between human behavior and human cortical responses in all visual cortical regions analyzed except three (hV4, mFus, and CoS, all marginally significant) (Fig. 4C). Early visual cortical regions were able to discriminate some categories, likely due to basic featural differences between categories, but not all categories (Fig. 4*C*; note the blue points in top left). These results suggest that the BOLD responses we measured in human visual cortex contain useful information for discriminating between different categories.

When discriminating natural images from feature-matched scrambled synths, cortical responses were unable to match the performance of human observers. Across several different image classes, we found that all observer models constructed using responses from each visual area were significantly less likely (V1: $t = -5.89$, $P < 0.001$; V2: $t = -6.25$, $P < 0.001$; V3: $t = -6.21$, $P < 0.001$; hV4: $t = -5.76$, $P < 0.001$; mFus: $t = -8.70$, $P < 0.001$; pFus: $t = -6.78$, $P < 0.001$; IOG: $t = -6.36$, $P < 0.001$; CoS: $t = -5.32$, $P < 0.001$; TOS: $t = -6.66$, $P < 0.001$; LO: $t = -5.93$, $P < 0.001$; VVC: $t = -7.76$, $P < 0.001$; PIT: $t = -5.24$, $P < 0.001$; VMV: $t = -6.53$, $P < 0.001$) to identify the natural object image compared to human observers (Fig. 4*D*). This suggests that visual cortical responses, across distinct functional and anatomical areas including early visual cortex (V1, V2, V3, hV4), ventral temporal cortex (mFus, pFus, CoS, VMV, VVC, PIT), and lateral occipital cortex (LO, IOG, TOS) do not preferentially represent the natural arrangement of object features relative to scrambled arrangements containing the same complex visual features, which suggests that representations in category-selective regions of the human visual cortex lack selectivity for natural feature arrangement.

Model of Macaque IT Cortex. These BOLD imaging data suggest that category-selective regions of the human visual cortex can discriminate different categories but are nonselective for natural feature arrangement. It is, however, possible that neural selectivity for natural feature arrangement is only observable at higher spatial (e.g., individual neurons) or temporal (e.g., individual spikes) resolutions. To attempt to address this concern in the absence of novel electrophysiological recordings, we employed an existing published dataset of macaque IT electrophysiological recordings as well as a dCNN-based model of those macaque IT neural sites.

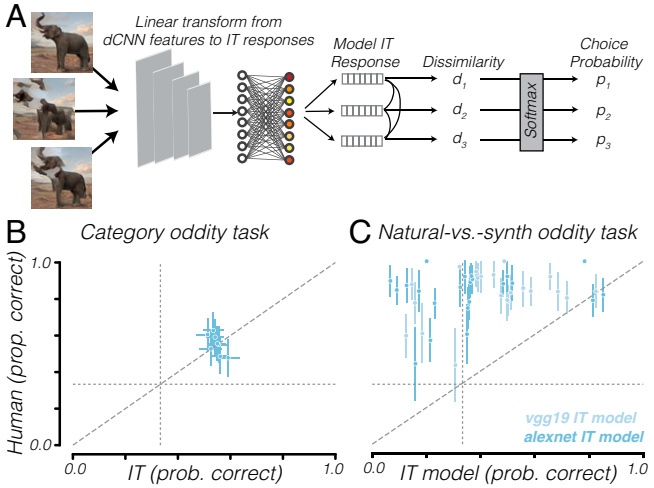

Fig. 5. IT observer model matches human performance at category discrimination but not natural-vs.-synth discrimination. (*A*) Schematic of IT observer model which uses 168 macaque IT multiunit responses to images as feature vectors and computes dissimilarity and choice probability similarly to the observer model in Fig. 1*B*. For the category oddity task, we examined responses measured from macaque IT cortex (71), whereas, for the natural–synth task, responses were simulated using a dCNN linearly fit to the dataset. (*B*) IT observer model vs. human performance on category oddity detection task, category by category. (*C*) IT observer model vs. human performance on natural–synth discrimination task, image by image. Diagonal dashed line is line of equality. Vertical and horizontal dotted lines represent chance level. Error bars indicate bootstrapped 95% CIs across trials. prob., probability; prop., proportion.

Using a dataset of macaque IT multiunit electrode recordings (71) as input to a cortical observer model, we found that, in a category oddity detection task, macaque IT responses matched human performance. That is, we used the spike rate (averaged across a temporal interval of 100 ms) recorded for each of 168 neural sites as a response vector which was used to compute oddity choice probability similarly to our other observer models. We compared this IT observer model's performance to human performance on a category oddity detection task using the same images presented to the macaques. We found that the human and IT observer model performance were not significantly different ($t = -0.235$, $P = 0.821$) (Fig. 5*B*), suggesting that macaque IT contains feature representations useful for discriminating between images of different categories.

To examine the ability of neurons in macaque IT cortex to perform the natural-vs.-synth oddity detection task, in the absence of electrophysiological recordings in response to our synths, we evaluated, instead, a dCNN-based model of IT neuronal response. This model was constructed by linearly transforming the feature space from a late convolutional layer of an Imagenet-trained dCNN to maximize predictivity of a population of 168 IT neural sites (71), an approach which has yielded state-of-the-art performance at predicting out-of-sample IT neural responses (48, 72). This IT model explains, on average across sites, 51.8% of the variance in neural response to held-out images (72). We used the responses of these model IT sites as inputs to an observer model constructed to perform the oddity detection task (Fig. 5*A*). On each trial, the observer model computed the response of the model IT sites to each of the three images, then computed dissimilarity and choice probability similarly to our other observer models. We evaluated two different dCNN models fit to IT response, one from the final convolutional layer of Alexnet and the other from VGG19 layer pool5.

We found that these dCNN-based models of IT sites were unable to match human performance discriminating natural from synthesized images. That is, in an oddity detection task where subjects saw one natural image and two synths with complex (pool4) features in a scrambled (1×1) arrangement, we found that humans were significantly more likely ($t = 9.46$, $P < 0.001$, $n = 87$) than a model of IT to pick the natural image across a wide variety of image classes (Fig. 5*C*). This finding suggests that the model IT population does not preferentially represent the natural arrangement of features compared to scrambled arrangements containing the same complex features.

Representational Geometry Analysis. The discrepancy in behavior on the natural-vs.-synth oddity detection task between human observers and dCNN models, category-selective regions of the human visual cortex, and macaque IT models suggests a misalignment in the underlying representational geometries which give rise to the observed behaviors. Therefore, we directly analyzed the representational spaces of dCNNs, category-selective visual areas, and model IT, and compared them to the perceptual representational space, which we inferred from behavioral responses in an independent perceptual task.

By estimating perceptual distances between images from an independent behavioral experiment, we found that the representations underlying human object perception must be selective for natural feature arrangement. Specifically, we presented two pairs of images on each trial and asked observers to choose which pair was more internally dissimilar (Fig. 6*A*) (73, 74). With the responses from this experiment pooled across all observers, we used a modified version of maximum likelihood difference scaling (MLDS) (73) to estimate the perceptual distances between pairs of images. Whereas the original MLDS method estimates the position of different stimuli along a single dimension to maximize the likelihood of the psychophysical responses (N free parameters for N stimuli), we directly estimated pairwise distances between each pair of stimuli ($\binom{N}{2}$ free parameters for N stimuli) to eliminate any assumptions about the dimensionality of the representational space. We visualized those distances by plotting the three images presented on a given trial of the oddity detection task in a triangle, such that the length of the edges corresponded to the pairwise perceptual distances, a visualization which we will refer to as a triangular distance plot (Fig. 6 *B–K*). We similarly visualized the representational geometries of the final convolutional layer of a dCNN model, human category-selective visual cortex, and a model of macaque IT neurons.

For the category discrimination task, we found that the representational geometry of human visual perception was well aligned with that of dCNNs, human category-selective visual cortex, and macaque IT neurons. The estimated perceptual distance between two images of two different categories significantly exceeded the estimated perceptual distance between two images of the same category ($t = 93.49$, $P < 0.001$) (Fig. 6*B*; note the narrow triangles). Similarly, for dCNN models (Fig. 6*C*), category-selective regions in the human visual cortex (Fig. 6*D*), and macaque IT cortex (Fig. 6*E*), the representational distance between images of different categories significantly exceeded the representational distance between images of the same category, albeit not quite as extremely as was found for human perception (Fig. 6*B*). This alignment of representational geometry (Fig. 6*F*) explains why human performance on the category oddity detection task was well matched by that of

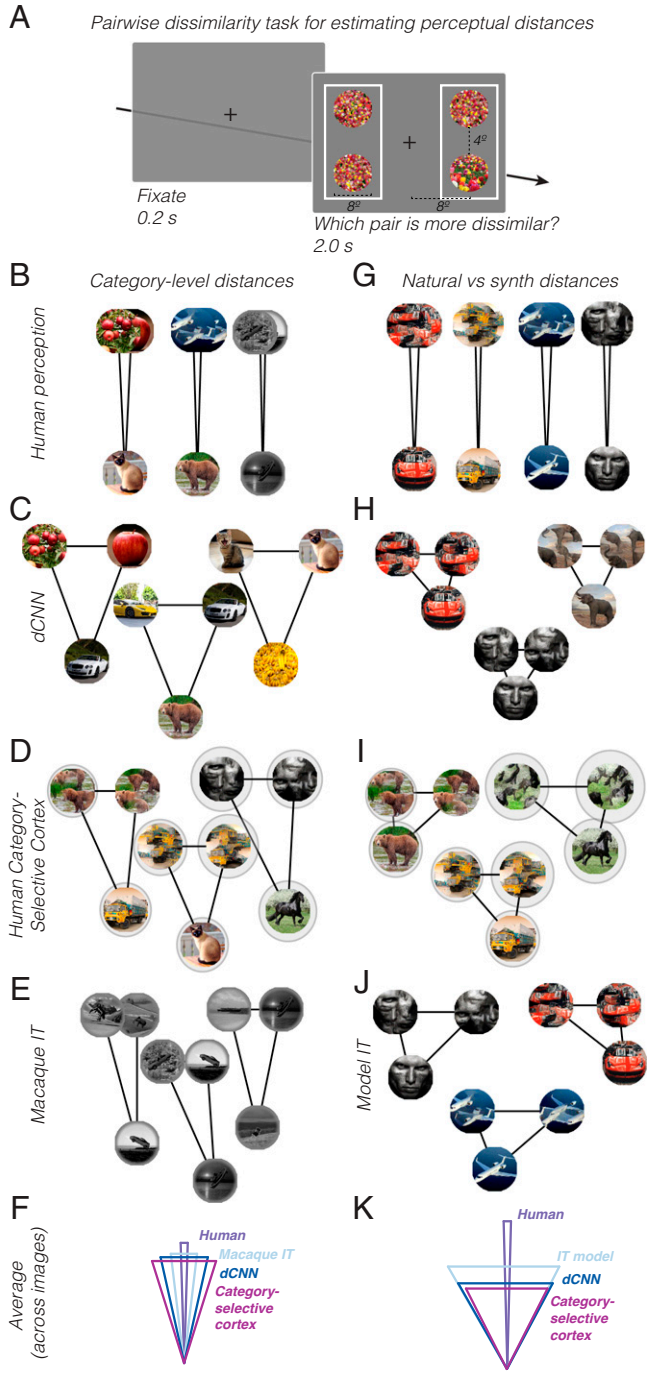

Fig. 6. Representational geometry misalignment between humans and model/cortical representations for natural vs. synth discrimination. (*A*) Pairwise dissimilarity judgment task. Subjects saw two pairs of images and reported which pair was more dissimilar, by key press. Responses from this task were used to estimate the relative perceptual distances between pairs of images. (*B*) Perceptual distances, estimated via modified MLDS, between images of different categories and images of the same category, for three example triplets, estimated across repeated presentations of each triplet. (*C*) The dCNN representational distances, averaged across dCNN models, between images of different categories and images of the same category, for three example triplets. (*D*) Cortical representational distances between images of different categories relative to images of same category, for three example triplets. Gray clouds around images represent split-half distance, that is, distance between neural response to each image on two halves of presentations. (*E*) Macaque IT representational distances between images of different categories and images of same category, for three example triplets. (*F*) Triangular distances averaged across all categories. (*G*–*K*) Same as *A*–*E* but for natural–synth representational distance relative to synth–synth distances for 1 × 1 pool4 synths.

dCNN observer models (Fig. 3*B*), category-selective regions of visual cortex (Fig. 4*B*), and macaque IT neurons (Fig. 5*B*).

However, for the natural-vs.-synth discrimination, we found a misalignment between the representational geometry of human visual perception and those of dCNNs, category-selective regions of the human visual cortex, and model IT. For human visual perception, we found that the estimated perceptual distance between the natural image and the synthesized images was significantly greater ($t = 69.25$, $P < 0.001$) than the perceptual distance between two different synths of the same class (Fig. 6*G*; note the narrow triangles), which reflects a selectivity for natural feature arrangement. In contrast, for all dCNN models we tested, we found that the representational distance between the natural image and the synthesized images was not significantly different ($t = -0.46$, $P = 0.65$, $n = 87$) from the representational distance between two different synthesized scrambled images (Fig. 6*H*; note the approximately equilateral triangles), suggesting that dCNN observer models are nonselective for natural feature arrangement of objects. Similarly, using BOLD responses from category-selective regions, we found that the representational distance between a natural image and a feature-matched synth was not significantly greater than the representational distance between two different synths of the same image class ($t = -1.52$, $P = 0.17$) (Fig. 6*I*), and, in a model of IT neurons, the representational distance between a natural image and a feature-matched synth was not significantly different ($t = 0.78$, $P = 0.44$) from the representational distance between two different synths (Fig. 6*J*). These findings demonstrate that representations in dCNN models, category-selective regions of the visual cortex, and a model of macaque IT neurons are nonselective for natural feature arrangement of objects and that the representational geometry of the visual cortex is therefore misaligned with that of human visual perception.

Control Analysis: Validating the Effect of Spatial Arrangement with Texture Stimuli. Our results thus far suggest that what differentiates human perception from visual cortical and dCNN representations is sensitivity to spatial arrangement of complex features. However, an alternate explanation is that the superior performance of human observers at discriminating natural from synthesized images might be driven by low-level artifacts in the synthesis process to which dCNN models and high-level visual cortex are insensitive. To control for this potential confound, we utilized a class of stimuli, visual textures, which are inherently defined by their invariance to the spatial arrangement of features. If artifacts in the image synthesis process were responsible for the superior performance of human observers, then we would expect that human observers should still be much more likely than dCNN or cortical observer models to identify the natural texture image as the odd one out. However, if the discrepancy in performance between humans and dCNN models is due to sensitivity to spatial arrangement as we hypothesized, then we would expect humans to be less able to accurately detect the natural texture image, resembling the performance of dCNN and visual cortical observer models.

We found that human observers' oddity detection performance for textures matched that of dCNN models, unlike for objects. We selected 12 natural texture images (e.g., bricks, rocks, grass, moss, and bark) with a relatively homogeneous spatial distribution of features and synthesized corresponding images which varied in their feature complexity and spatial arrangement. We compared the performance of human observers with five different dCNN observer models as we varied both the feature complexity and spatial arrangement of the

synthesized images of textures. We found that human observers' likelihood of identifying the natural image was not significantly different (linear mixed effects model: $b = 0.027$, SE = 0.014, $P = 0.066$, 95% CI = [−0.002, 0.055]) from the dCNN models (*SI Appendix*, Fig. S1 *A* and *B*). Using a representational geometry analysis, we found that the perceptual distance between natural texture images and synths only marginally exceeded ($t = 3.38$, $P = 0.006$) the perceptual distance between two different synths (*SI Appendix*, Fig. S1*C*). In comparison to objects, the perceptual representation of textures was significantly less selective for natural feature arrangement ($t = -13.96$, $P < 0.001$). As expected, the representational geometry of dCNNs (*SI Appendix*, Fig. S1*D*) and human category-selective cortex (*SI Appendix*, Fig. S1*E*) for textures also reflected a similar nonselectivity for the natural arrangement of features. In sum, these results suggest that the representations in the human visual cortex and dCNN models can better account for human perception of textures, which are invariant to feature arrangement, than of objects.

Do Neural and dCNN Representations Contain Information about Spatial Arrangement? To reconcile the discrepancy between neural representations and perception with regards to selectivity for natural feature arrangement, we sought to find a transformation of the cortical representation that might better approximate human perception. To quantify the ability of a particular representational space to distinguish natural objects from synths with scrambled matching features, we developed a natural image selectivity index that measures the degree to which the representational distance between the natural image and the synths exceeds the distance between two different synths.

$$\text{Selectivity Index} = \frac{d_{natural,synth} - d_{synth1,synth2}}{d_{natural,synth} + d_{synth1,synth2}}.$$

We found that an unweighted readout of the cortical representation yielded a natural image selectivity index that was not significantly different from zero (Fig. 7*A*, purple bars) in nearly all category-selective visual areas (mFus: $t = 1.21$, $P = 0.27$; pFus: $t = 4.66$, $P = 0.003$; IOG: $t = 1.24$, $P = 0.26$; CoS: $t = 2.03$, $P = 0.08$; TOS: $t = 2.45$, $P = 0.05$; LO: $t = 0.97$, $P = 0.37$; VVC: $t = 2.09$, $P = 0.08$; PIT: $t = -0.13$, P = 0.90; VMV: $t = 3.10$, $P = 0.02$), with the marginal exception of pFus, TOS, and VMV. This contrasts sharply with the selectivity of perception (Fig. 7*A*, dashed line), in line with our finding that the cortical observer model was unable to explain human behavioral performance (Fig. 4*D*). We hypothesized that a linear transformation of cortical responses might yield a representational space that is more selective for the natural arrangement of object features. To test this hypothesis, using gradient descent, we sought to find a linear weighting of cortical responses which would yield a representation that was as selective for the natural arrangement of features as human perception.

We first tested a linear transform of visual cortical responses that generalized across image classes but found that such a readout failed to increase selectivity. We used gradient descent to find a linear weighting of cortical responses that would maximize the natural feature arrangement selectivity index. To ensure that this transform would generalize, we cross-validated across image classes. That is, we fit the weights to maximize the natural image selectivity for all but one image class and then evaluated the selectivity of the weighted representation on the held-out image class. This procedure failed to yield a representation which was significantly more selective for natural feature

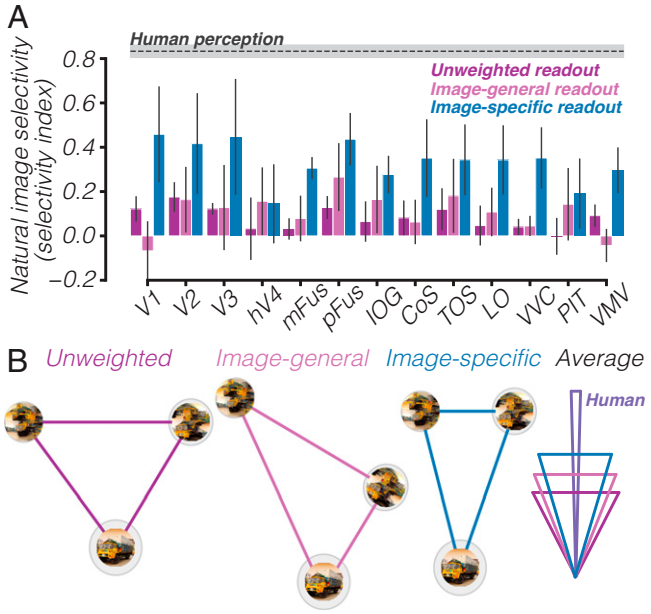

Fig. 7. Evidence that information about natural feature arrangement can be read out from the visual cortex. (*A*) Natural image selectivity index across all analyzed cortical regions, for unweighted readout (purple), image-general readout, where weights were fit on one image class and evaluated on held-out image class (magenta), and image-specific readout, where separate weights were fit for each image class and evaluated on held-out presentations (blue). Gray dashed line shows natural image selectivity of human perception. Error bars indicate 95% CIs across image classes. (*B*) Representational geometry for one representative area (LO) for one example image class, comparing between (*Left*) unweighted, (*Left Center*) image-general, and (*Right Center*) image-specific readouts. (*Right*) The representational geometry, averaged across all image classes.

arrangement than the unweighted readout in all regions of the visual cortex (V1: $t = 2.05$, $P = 0.09$; V2: $t = 0.23$, $P = 0.82$; V3: $t = 0.00$, $P = 1.00$; V4: $t = -1.02$, $P = 0.35$; mFus: $t = -0.75$, $P = 0.48$; pFus: $t = -1.72$, $P = 0.14$; IOG: $t = -1.09$, $P = 0.32$; CoS: $t = 0.56$, $P = 0.59$; TOS: $t = -0.57$, $P = 0.59$; LO: $t = -0.91$, $P = 0.40$; VVC: $t = -0.69$, $P = 0.52$; PIT: $t = -1.94$, $P = 0.10$; VMV: $t = 1.84$, $P = 0.12$) (Fig. 7*A*, magenta bars).

We next tested whether fitting a separate linear transform for each image class could increase the natural image selectivity of the representation. This would suggest that the representation contains enough information to extract the natural feature arrangement, even if it is not generalizable across image classes. To do so, we used gradient descent to find, for each image class, a weighting of cortical responses that would maximize the natural image selectivity index for that particular image class. Therefore, for each subject, within each brain area, we fit 1,000 weights (100 voxels × 10 image classes). Since each image was presented multiple times, we cross-validated across presentations, such that weights were fit based on trial-averaged responses from 90% of the presentations, and selectivity was assessed on the held-out 10% of image presentations.

We found that this image-specific weighting of cortical responses significantly increased the natural feature arrangement selectivity of the neural representation (V1: $t = -3.03$, $P = 0.02$; V2: $t = -1.94$, $P = 0.10$; V3: $t = -2.45$, $P = 0.05$; V4: $t = -2.38$, $P = 0.06$; mFus: $t = -12.94$, $P < 0.001$; pFus: $t = -5.94$, $P = 0.001$; IOG: $t = -3.63$, $P = 0.01$; CoS: $t = -3.41$, $P = 0.01$; TOS: $t = -2.74$, $P = 0.03$; LO: $t = -4.85$, $P = 0.004$; VVC: $t = -4.74$, $P = 0.003$; PIT: $t = -2.81$, $P = 0.03$; VMV: $t = -4.13$, $P = 0.01$) (Fig. 7*A*, blue bars) in all cortical areas, although it was still unable to reach

the level of selectivity observed behaviorally (V1: $t = -2.99$, $P = 0.024$; V2: $t = -3.00$, $P = 0.024$; V3: $t = -2.53$, $P = 0.045$; V4: $t = -7.34$, $P < 0.001$; mFus: $t = -18.81$, $P < 0.001$; pFus: $t = -6.79$, $P < 0.001$; IOG: $t = -11.47$, $P < 0.001$; CoS: $t = -4.65$, $P = 0.003$; TOS: $t = -5.80$, $P = 0.001$; LO: $t = -5.61$, $P = 0.001$; VVC: $t = -6.40$, $P = 0.001$; PIT: $t = -7.18$, $P < 0.001$; VMV: $t = -8.55$, $P < 0.001$). This form of readout, however, requires prior experience with every individual image to learn an image-specific linear transform, so it is unlikely to be a plausible mechanism by which visual cortical responses could directly support perception. Rather, it demonstrates that the information about natural feature arrangement can be found in these visual areas but would likely require further untangling to yield a representation which is more predictive of behavior.

We corroborated these results using a support vector machine (SVM) classifier with a linear kernel trained to classify images as natural or synthesized. When trained using either cortical responses (*SI Appendix*, Fig. S2*A*) or dCNN features (*SI Appendix*, Fig. S2*B*), the SVM classifier was only able to classify images accurately if those image classes were included in its training set.

It is possible that the final layer of dCNN models, which transforms the feature representation into class probabilities for image categorization, is more selective for natural object feature arrangement than the convolutional layers. To test this, we analyzed the representational geometry of the last fully connected layer in each dCNN model in comparison to that of the final convolutional layer. We found that, in three out of four dCNN models, the last fully connected layer was marginally more selective for natural feature arrangement compared to the last convolutional layer ($t = -2.98$, $P = 0.058$) (*SI Appendix*, Fig. S6*A*), although it was still significantly less selective for natural feature arrangement than human observers ($t = -14.90$, $P < 0.001$). This demonstrates the plausibility of a nonlinear readout to transform the feature representation in dCNNs or the visual cortex to yield a representation which better matches perceptual selectivity for natural feature arrangements of objects.

Discussion

Through the measurement and analysis of human behavior, cortical responses, and dCNN features, we sought to characterize the visual representations which enable the perception of natural objects. We found that human visual perception is sensitive to the complexity of features in objects and selective for the natural arrangement of object features. In contrast, we found that both dCNN features and visual cortical responses were relatively poor at distinguishing natural images from synthesized scrambles containing complex features. A model of macaque IT neurons was also similarly insensitive to the arrangement of features in object images. This insensitivity was not due to a lack of featural representations but rather due to an insensitivity to spatial arrangement, as demonstrated by the evidence that representations in dCNNs, category-selective regions of the human visual cortex, and macaque IT matched human performance in both a category oddity task and a texture oddity task, which did not require selectivity for feature arrangement. Thus, we concluded that both the human visual cortex and dCNN models do not preferentially represent natural object images compared to scrambled images of complex object features and are therefore unable to directly account for the human perceptual ability to discriminate natural from synthesized images of objects. This suggests that further computation beyond feed-forward ventral visual cortical response is required to support human object perception (75), although the complex visual features represented are a useful intermediate representation. To confirm this, we demonstrated that the information necessary to match perceptual selectivity for natural images is decodable from category-selective regions of the human visual cortex, although it requires a specialized image-specific readout. Taken in sum, our results suggest that the representations found in the human visual cortex, macaque IT cortex, and Imagenet-trained dCNNs encode the complex visual features that make up objects, although they do not preferentially encode the natural arrangement of features which defines an object.

Although the image synthesis technique that we employed allowed us to separately control the complexity and spatial arrangement of visual features in synthesized images, it is likely that feature complexity and spatial arrangement are not fully independent dimensions. That is, complex visual features are composed of particular arrangements of simpler features. This is best exemplified by the recent finding that a dCNN with random filters at multiple spatial scales can be used to synthesize textures which are perceptually similar to the original and comparable to the quality of textures synthesized by a trained model (6, 76). Nonetheless, it is useful to artificially vary the spatial arrangement of features at different levels of complexity to assess the contributions of each of these to perception. This utility is best demonstrated by the contrast between object and texture perception: Whereas human observers' perception of textures was sensitive to the complexity of visual features but not the spatial arrangement of those complex features, object perception was sensitive to both the complexity and arrangement of visual features. Further, this disentangling of feature complexity and spatial arrangement is justified by the finding that, at a particular level of feature complexity (pool4), there was a mismatch between human perception and cortical representations in terms of selectivity for natural images relative to synthesized images.

To demonstrate the lack of cortical selectivity for natural feature arrangement, we used two convergent pieces of cortical evidence: BOLD imaging of the human visual cortex and a model of macaque IT neurons. The BOLD imaging data were limited in their spatial and temporal resolution but offered broad coverage of many visual areas across the occipital and temporal lobes. To address the possibility that selectivity for natural feature arrangement might only be observable at higher resolutions, we utilized a well-validated model of macaque IT neurons. This method, too, had its limitations, as the model does not explain all the variance in IT neural responses, and it is possible that the unexplained variance might account for the perceptual selectivity for natural feature arrangement. However, even state-of-the-art techniques can only measure from a few thousand neurons in a localized region of the brain, which would leave open the possibility that selectivity for natural feature arrangements might be found when considering a larger, more representative sample of neurons. Therefore, given that both of our approaches, one with wide spatial coverage and another which simulates individual neuronal responses, yielded convergent results, we conclude that all our available evidence suggests there is a mismatch between human perception and representations in category selective cortex.

It is possible that this mismatch between human perception and cortical representations is due to subjects in the neuroimaging experiment passively viewing the images while performing a fixation task, in contrast to subjects in the behavioral experiment who were actively attending to the images. This possibility might imply that feedback, in the form of attention or other

top-down signals, transforms visual cortical representations to facilitate the discrimination of natural from synthesized images. Therefore, our results should only be taken to apply to the feedforward cortical representations generated while passively viewing images. However, given that these passively generated feedforward representations were sufficient to match behavioral performance in the category oddity task, it is nonetheless noteworthy that feedforward responses do not preferentially represent natural images relative to synthesized scrambled images containing similar features.

Our results contribute to a large body of research about the texture-like encoding of peripheral vision (10, 18, 19). Despite the stimuli in our experiments being presented in the visual periphery, human subjects were highly sensitive to the spatial arrangement of features for object-like stimuli but not for texture-like stimuli, in line with recent findings (14). We extend these findings to demonstrate that visual cortical representations of peripherally presented objects are not selective for the natural spatial arrangement of visual features. It is possible that, had we measured cortical responses to foveally presented stimuli, we might have found a greater degree of selectivity for natural feature arrangement. However, the stimuli from the macaque IT dataset were presented at the center of gaze, yet we found that the model IT representation was insensitive to natural feature arrangement. Further, we presented the stimuli peripherally in the behavioral experiments while enforcing fixation (*SI Appendix*, Fig. S3) and nonetheless observed selectivity for natural feature arrangement, so it is unlikely that our results could be explained by the texture-like encoding of peripheral vision alone.

These results suggest a possible cortical mechanism to explain a series of discrepant findings in the scene perception literature regarding why human observers cannot distinguish feature-matched synths (10) but can easily distinguish natural images of objects from feature-matched synths (14). An influential study (10) demonstrated that matching first- and second-order statistics within spatial pooling regions, whose size matched V2 receptive fields, results in metameric images. However, subsequent work (14, 15) has shown that this metamerism only holds when comparing two synthesized samples but breaks down when one of the samples is the original image. It is particularly the case that human subjects can easily discriminate natural images from synths with scrambled features for objects or scenes more than for textures (14). If the synths are generated to match the locally pooled features of the natural image, then why might humans be able to discriminate the natural image from a synth but unable to discriminate two different synths? Our behavioral results corroborate the findings of ref. 14, suggesting that image content—that is, whether an image contains an object or a texture—matters beyond just the size of spatial pooling regions in terms of whether a natural image is perceptually distinct from synthesized counterparts. Our BOLD imaging data corroborate the findings of refs. 10 and 21, that the ventral stream contains texture-like representations, but also suggests a possible cortical mechanism by which this texture-like representation can support the enhanced discriminability of natural objects from synths, via a specialized object-specific readout.

In the domain of face perception, prior research has sought to distinguish selectivity for complex features from selectivity for the arrangement of those features. There is evidence suggesting that face-selective visual areas in the ventral temporal cortex are sensitive to the spatial arrangement of facial features, as demonstrated by the finding that the mean response of all voxels in the fusiform face area is greater to intact faces than faces containing scrambled facial features (60), as well as the finding that a linear classifier can decode intact vs. scrambled faces (61), which might seem to contradict our findings. However, these prior findings relied on handpicked features or grid-based scrambling approaches. Due to the limited number of stimuli in our sample, we did not specifically examine selectivity for spatial arrangement of facial features, so we cannot make any claims about whether our approach using deep image synthesis would confirm or contradict prior studies using handpicked features. Regardless, given the degree to which face-selective cortical populations appear to be highly functionally specialized and anatomically segregated compared to other object-selective populations in VTC and LO, it is certainly possible that the neural representation of faces might be explicitly selective for the natural arrangement of facial features, while most other objects are represented by a collection of disjointed complex visual features.

Our results contribute to a long-standing debate about whether the perception of objects is holistic or featural. Behavioral effects, such as the Thatcher effect (77), in which subjects are relatively insensitive to local feature rotations in an upside-down face, or the finding that humans are better at identifying facial features when presented in context than in isolation (78, 79), have given rise to the view that objects must be represented holistically, not as an independent set of features. In the present study, we assess holistic perception as the degree to which the whole object is perceptually distinct from the scrambled features of the object. Using the oddity detection task, we found that the natural image is far more perceptually distinct from a synthesized scrambled image than two different scrambled images are from each other, suggesting that humans perceive objects holistically. However, cortically, we found that natural objects are not represented more distinctly from synths than two different synths are from one another, suggesting a nonholistic representation. These results suggest that holistic perception of objects may arise from a featural cortical representation.

Our findings build upon recent research demonstrating that Imagenet-trained dCNN models are texture-biased, that is, more likely to make use of texture than shape information when classifying images, in contrast to humans who are shape biased (53, 54, 56, 80). However, while these prior studies were taken as evidence that dCNNs are a poor model of visual representations in the human brain, our results demonstrate that representations in the visual cortex are as texture-like as those of dCNNs, and, therefore, both dCNNs and category-selective regions of the visual cortex similarly deviate from human behavior. When texture and shape cues are artificially made to conflict, either by using silhouetting (54) or by using neural style transfer (53), Imagenet-trained dCNN models are more likely to classify the conflicting images according to the texture than the shape label, unlike humans. It is possible that the texture bias of dCNNs is driven by the readout rather than the feature representations themselves. Our results demonstrate that dCNNs' texture bias is driven by the lack of selectivity for natural spatial arrangement in the feature representation itself, not by a texture-biased readout. We note that the term shape, while related to the concept of feature arrangement, is often used to refer to external contour. While our image synthesis technique does not specifically target external contours, it does disrupt them along with any other natural arrangement of features. Finally, our findings examining cortical representations of objects suggest a reinterpretation of the texture bias results (53, 54). In contrast to the suggestion that the texture bias of dCNN models makes them flawed as models of human vision,

we found that the human visual cortex similarly contains texture-like representations of objects. We can thus speculate that this texture-like representation, which is nonselective for natural feature arrangements of objects, might be useful for the perception of objects. This might seemingly contradict the finding that texture-biased dCNNs are more susceptible to image distortions than their shape-biased counterparts (53). However, our results suggest the possibility that texture-like representations need not necessarily be less robust, if coupled with a sufficiently sophisticated readout.

What potential advantage is there to a representation that encodes complex visual features but does not prioritize the natural arrangement of these features? Classic theories of object vision have posited that objects can be identified by the arrangement of more simple building blocks or primitives (43). Our findings suggest that textures are one of these basic building blocks that are encoded by category selective cortex. Category selectivity of visual cortical areas then comes from visual features that are informative for particular category judgements. For example, parahippocampal place area, although known for selectivity for scenes and places (81, 82), is also highly active for more simple rectilinear features that are building blocks for that category (83). Our findings further show that, despite not being selective for the natural arrangement of visual features, that information has not been lost in the cortical representation. That is, a classification analysis could decode (on an item-by-item basis) the natural spatial arrangement. If sensory representations were, instead, specific for a particular arrangement of complex visual features, this might preclude the possibility of learning new arrangements of features for novel object categories. The implication is that it is beneficial to have a feature representation in high-level visual cortex that is nonselective for natural spatial arrangement, because it might allow for more rapid, robust transfer learning, the repurposing of learned features for novel objects or tasks. This is certainly true of Imagenet-trained dCNN models, whose learned feature space can be easily repurposed for new object categories or even for other tasks (84, 85), simply by learning a new linear readout. Taken together, our results suggest that cortical visual responses in category-selective regions represent not objects, per se, but a basis set of complex visual features that can be infinitely transformed into representations of the myriad objects and scenes that we encounter in our visual environments.

Materials and Methods

Behavioral Methods.

Observers. Eighty-seven observers, naive to the goals of this study, performed 100 trials of the natural–synth oddity detection behavioral experiment, 85 observers performed 50 trials of the category oddity experiment, and 110 observers performed 100 trials of the dissimilarity judgment experiment, on Amazon Mechanical Turk. Subjects were eligible to participate if their human intelligence task (HIT) approval rate (percentage of completed HITs that were approved by prior requesters) exceeded 75%. Subjects' data were excluded only if their performance on trivial catch trials failed to exceed 40% accuracy (chance level performance was 33%). To validate the online findings, in-laboratory behavioral data were collected from two observers, where each observer performed at least 5,000 trials. For the neuroimaging experiment, seven observers (four female, three male, mean age 28 y, age range 23 y to 37 y), six naive to the goals of the study, participated as subjects in two 1-h scans at the Stanford Center for Neurobiological Imaging. All protocols were approved beforehand by the Institutional Review Board for research on human subjects at Stanford University, and all observers gave informed consent prior to the start of the experiment by signing a form.

Stimulus generation. The stimuli used in all experiments were either natural images of real objects/textures or synthetically generated through an iterative optimization procedure ("synths") designed to match features from a target image. To synthesize images, we passed a target natural image into an Imagenet-trained VGG-19 dCNN model (62) and extracted the activations from three intermediate layers (pool1, pool2, and pool4) (Fig. 1*A*). We calculated a spatially constrained Gramian matrix, that is, the inner product between every pair of activation maps, of each layer's activations, which allowed us to spatially pool features over image subregions of predefined sizes. Finally, we iteratively updated the pixels of a random white noise image to minimize the mean-squared error between the spatially weighted Gramians of the output image and the target image. Thus, by varying the size of the spatial pooling regions, we could vary the degree to which feature arrangement was spatially constrained, and, by varying which layers' features were included in the loss function of the optimization, we could control the complexity of visual features in the synthesized image. See *SI Appendix, Extended Methods* for more detail.

Experimental design: Oddity detection task. On each trial (Fig. 2*A*), observers were concurrently presented with three images for up to 2 s and asked to respond, by key press (up, left, or right), anytime during the 2-s interval, indicating which image appeared most different from the others. In the natural–synth task, one image was a natural image, and the other two were synths, matched to the features of the natural image, at a particular dCNN layer and spatial constraint level. We performed the natural–synth experiment both on Amazon Mechanical Turk ($n = 87$ subjects, 6,165 trials) and in the laboratory ($n = 2$ subjects, 5,000 trials each). In the category oddity task, two images contained objects of the same category, and the third image contained an object from a different category. This experiment was performed on Amazon Mechanical Turk ($n = 85$ subjects, 3,448 trials). See *SI Appendix, Extended Methods* for more detail.

Experimental design: Pairwise dissimilarity task. From an independent set of 110 observers, we measured relative dissimilarity judgments of pairs of images. On each trial, subjects were concurrently presented with four images, grouped into two pairs, and were asked to indicate, with a left or right key press, which of the two pairs was more dissimilar. See *SI Appendix, Extended Methods* for more detail.

Estimating perceptual distances. Using behavioral responses from the pairwise dissimilarity judgment task, we estimated the perceptual distances between pairs of images. We used a modified version of the maximum-likelihood difference scaling model (73) to estimate the perceptual distances between pairs of images as free parameters in an optimization procedure designed to maximize the likelihood of the observed behavioral responses. See *SI Appendix, Extended Methods* for more detail.

dCNN Methods. We modeled task performance using features extracted from a dCNN. We tested the representational space of five different Imagenet-trained dCNNs: VGG-19 (62), CORnet-Z (63), VGG-16 (62), ResNet-18 (64), and AlexNet (46). On each trial, our model extracted a feature vector from the last convolutional layer of the dCNN for each image presented (Fig. 3*A*). Next, we computed the Pearson distance between the features of each pair of images, and, for each image, calculated its dissimilarity as the mean Pearson distance between that image and each of the other two images. Finally, the model converted these dissimilarities into choice probabilities using a Softmax transform. See *SI Appendix, Extended Methods* for more detail.

To analyze the representational space learned by Imagenet-trained dCNN models, we employed a representational similarity analysis to compute the Pearson correlation distance between the activations from the last convolutional layer in response to each pair of images (86). We then used these representational distances to determine the selectivity of this feature space for natural feature arrangement, computed as

$$\frac{d_{natural,synth} - d_{synth1,synth2}}{d_{natural,synth} + d_{synth1,synth2}}.$$

Modeling IT Neurons. To estimate object selectivity of neurons in IT cortex, we modeled the response of each IT neural site (71) as a linear combination of dCNN units (48, 51, 87, 88). Using this population of model IT neurons, we modeled oddity detection task performance and calculated the representational

selectivity for natural feature arrangement. See *SI Appendix, Extended Methods* for more detail.

Neuroimaging methods. We used BOLD imaging (89) to measure cortical responses to visually presented images, both natural and synthesized (1 × 1 pool4), from 10 different categories. Images subtended 12°. We defined cortical areas using population receptive field mapping to identify retinotopic areas in early visual cortex (68, 90–92), in addition to an atlas-based approach to identify anatomically defined areas (70) and a functional localizer to identify category-selective regions (69). Using a generalized linear model (66, 67), we extracted trial-averaged responses to individual images. We used these responses as input to observer models to perform the task as well as to compute the selectivity of neural representations for natural feature arrangement. See *SI Appendix, Neuroimaging Methods* for more details.

Data Availability. BOLD imaging data have been deposited in Open Science Framework (https://osf.io/gpx7y/).

ACKNOWLEDGMENTS. We acknowledge the support of Research to Prevent Blindness and Lions Clubs International Foundation and the Hellman Fellows Fund to J.L.G. This work was supported by the Stanford Center for Cognitive and Neurobiological Imaging. We would like to thank Tyler Bonnen, Nathan Kong, Dan Yamins, Kalanit Grill-Spector, Tony Norcia, Josh Wilson, Josh Ryu, Austin Kuo, and Jiwon Yeon for their feedback on previous versions of this manuscript.

1. T. Young II, The Bakerian lecture. On the theory of light and colours. *Philos. Trans. R. Soc. Lond.* **92**, 12–48 (1802).
2. B. A. Wandell, *Foundations of Vision* (Oxford University Press).
3. B. Julesz, Visual pattern discrimination. *Ire. T. Inform. Theor.* **8**, 84–92 (1962).
4. B. Julesz, Textons, the elements of texture perception, and their interactions. *Nature* **290**, 91–97 (1981).
5. J. Portilla, E. P. Simoncelli, A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis.* **40**, 49–70 (2000).
6. L. Gatys, A. S. Ecker, M. Bethge, Texture synthesis using convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **28**, 262–270 (2015).
7. L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, E. Shechtman, "Controlling perceptual factors in neural style transfer" in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Institute of Electrical and Electronics Engineers, 2017), pp. 3730–3738.
8. J. Feather, A. Durango, R. Gonzalez, J. McDermott, "Metamers of neural networks reveal divergence from human perceptual systems" in *Advances in Neural Information Processing Systems*, H. Wallach *et al.*, Eds. (Curran Associates, 2019), pp. 10078–10089.
9. T. S. A. Wallis *et al.*, A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *J. Vis.* **17**, 5 (2017).
10. J. Freeman, E. P. Simoncelli, Metamers of the ventral stream. *Nat. Neurosci.* **14**, 1195–1201 (2011).
11. A. C. Bovik, M. Clark, W. S. Geisler, Multichannel texture analysis using localized spatial filters. *Ieee T Pattern Anal* **12**, 55–73 (1990).
12. B. J. Balas, Texture synthesis and perception: Using computational models to study texture representations in the human visual system. *Vision Res.* **46**, 299–309 (2006).
13. T. S. A. Wallis, M. Bethge, F. A. Wichmann, Testing models of peripheral encoding using metamerism in an oddity paradigm. *J. Vis.* **16**, 4 (2016).
14. T. S. Wallis *et al.*, Image content is more important than Bouma's Law for scene metamers. *eLife* **8**, e42512 (2019).
15. A. Deza, A. Jonnalagadda, M. Eckstein, Towards metamerism via foveated style transfer. arXiv [Preprint] (2017). https://doi.org/10.48550/arXiv.1705.10041 (Accessed 7 April 2022).
16. B. Balas, L. Nakano, R. Rosenholtz, A summary-statistic representation in peripheral vision explains visual crowding. *J. Vis.* **9**, 13 (2009).
17. R. Rosenholtz, J. Huang, A. Raj, B. J. Balas, L. Ilie, A summary statistic representation in peripheral vision explains visual search. *J. Vis.* **12**, 14 (2012).
18. R. Rosenholtz, Capabilities and limitations of peripheral vision. *Annu. Rev. Vis. Sci.* **2**, 437–457 (2016).
19. D. G. Pelli, K. A. Tillman, The uncrowded window of object recognition. *Nat. Neurosci.* **11**, 1129–1135 (2008).
20. D. M. Levi, Crowding–An essential bottleneck for object recognition: A mini-review. *Vision Res.* **48**, 635–654 (2008).
21. J. Freeman, C. M. Ziemba, D. J. Heeger, E. P. Simoncelli, J. A. Movshon, A functional and perceptual signature of the second visual area in primates. *Nat. Neurosci.* **16**, 974–981 (2013).
22. C. M. Ziemba, J. Freeman, J. A. Movshon, E. P. Simoncelli, Selectivity and tolerance for visual texture in macaque V2. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E3140–E3149 (2016).
23. C. M. Ziemba, J. Freeman, E. P. Simoncelli, J. A. Movshon, Contextual modulation of sensitivity to naturalistic image structure in macaque V2. *J. Neurophysiol.* **120**, 409–420 (2018).
24. G. Okazawa, S. Tajima, H. Komatsu, Image statistics underlying natural texture selectivity of neurons in macaque V4. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E351–E360 (2015).
25. G. Okazawa, S. Tajima, H. Komatsu, Gradual development of visual texture-selective properties between macaque areas V2 and V4. *Cereb. Cortex* **27**, 4867–4880 (2017).
26. R. A. Epstein, C. I. Baker, Scene perception in the human brain. *Annu. Rev. Vis. Sci.* **5**, 373–397 (2019).
27. M. D. Lescroart, J. L. Gallant, Human scene-selective areas represent 3D configurations of surfaces. *Neuron* **101**, 178–192.e7 (2019).
28. N. C. Rust, J. J. DiCarlo, Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *J. Neurosci.* **30**, 12978–12995 (2010).
29. J. V. Haxby *et al.*, Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).
30. J. J. DiCarlo, D. Zoccolan, N. C. Rust, How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
31. K. Grill-Spector, K. S. Weiner, The functional architecture of the ventral temporal cortex and its role in categorization. *Nat. Rev. Neurosci.* **15**, 536–548 (2014).
32. N. Kanwisher, J. McDermott, M. M. Chun, The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).
33. R. Epstein, A. Harris, D. Stanley, N. Kanwisher, The parahippocampal place area: Recognition, navigation, or encoding? *Neuron* **23**, 115–125 (1999).
34. L. Cohen *et al.*, The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain* **123**, 291–307 (2000).
35. P. E. Downing, Y. Jiang, M. Shuman, N. Kanwisher, A cortical area selective for visual processing of the human body. *Science* **293**, 2470–2473 (2001).
36. T. Serre, A. Oliva, T. Poggio, A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 6424–6429 (2007).
37. N. Kriegeskorte, Relating population-code representations between man, monkey, and computational models. *Front. Neurosci.* **3**, 363–373 (2009).
38. A. Pasupathy, C. E. Connor, Population coding of shape in area V4. *Nat. Neurosci.* **5**, 1332–1338 (2002).
39. S. O. Murray, P. Schrater, D. Kersten, Perceptual grouping and the interactions between visual cortical areas. *Neural Netw.* **17**, 695–705 (2004).
40. C. Cadieu *et al.*, A model of V4 shape selectivity and invariance. *J. Neurophysiol.* **98**, 1733–1750 (2007).
41. A. Pasupathy, T. Kim, D. V. Popovkina, Object shape and surface properties are jointly encoded in mid-level ventral visual cortex. *Curr. Opin. Neurobiol.* **58**, 199–208 (2019).
42. E. Margalit, I. Biederman, B. S. Tjan, M. P. Shah, What is actually affected by the scrambling of objects when localizing the lateral occipital complex? *J. Cogn. Neurosci.* **29**, 1595–1604 (2017).
43. I. Biederman, Recognition-by-components: A theory of human image understanding. *Psychol. Rev.* **94**, 115–147 (1987).
44. K. Fukushima, Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36**, 193–202 (1980).
45. M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025 (1999).
46. A. Krizhevsky, I. Sutskever, G. Hinton, "ImageNet classification with deep convolutional neural networks" in *Advances in Neural Information Processing Systems*, F. Pereira *et al.*, Eds. (Curran Associates, 2012).
47. J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database" in 2009 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Institute of Electrical and Electronics Engineers, 2009), pp. 248–255.
48. D. L. K. Yamins *et al.*, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
49. S.-M. Khaligh-Razavi, N. Kriegeskorte, Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
50. U. Güçlü, M. A. J. van Gerven, Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).
51. M. Schrimpf *et al.*, Brain-score: Which artificial neural network for object recognition is most brain-like? bioRxiv [Preprint] (2020). https://doi.org/10.1101/407007 (Accessed 7 April 2022).
52. C. Zhuang *et al.*, Unsupervised neural network models of the ventral visual stream. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2014196118 (2021).
53. R. Geirhos *et al.*, ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv [Preprint] (2018). https://doi.org/10.48550/arXiv.1811.12231 (Accessed 7 April 2022).
54. N. Baker, H. Lu, G. Erlikhman, P. J. Kellman, Deep convolutional networks do not classify based on global object shape. *PLoS Comput. Biol.* **14**, e1006613 (2018).
55. A. A. Zeman, J. B. Ritchie, S. Bracci, H. Op de Beeck, Orthogonal representations of object shape and category in deep convolutional neural networks and human visual cortex. *Sci. Rep.* **10**, 2453 (2020).
56. K. L. Hermann, T. Chen, S. Kornblith, The origins and prevalence of texture bias in convolutional neural networks in *Advances in Neural Information Processing Systems*, H. LaRochelle *et al.*, Eds. (Curran Associates, 2019), pp. 19000–19015.
57. B. Long, C.-P. Yu, T. Konkle, Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E9015–E9024 (2018).
58. K. Grill-Spector *et al.*, A sequence of object-processing stages revealed by fMRI in the human occipital lobe. *Hum. Brain Mapp.* **6**, 316–328 (1998).
59. Y. Lerner, T. Hendler, D. Ben-Bashat, M. Harel, R. Malach, A hierarchical axis of object processing stages in the human visual cortex. *Cereb. Cortex* **11**, 287–297 (2001).
60. J. Liu, A. Harris, N. Kanwisher, Perception of face parts and face configurations: An FMRI study. *J. Cogn. Neurosci.* **22**, 203–211 (2010).
61. J. Zhang, J. Liu, Y. Xu, Neural decoding reveals impaired face configural processing in the right fusiform face area of individuals with developmental prosopagnosia. *J. Neurosci.* **35**, 1539–1548 (2015).
62. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv [Preprint] (2014). https://doi.org/10.48550/arXiv.1409.1556 (Accessed 7 April 2022).
63. J. Kubilius *et al.*, CORnet: Modeling the neural mechanisms of core object recognition. bioRxiv [Preprint] (2018). https://doi.org/10.1101/408385 (Accessed 7 April 2022).
64. K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition" in *2016 Conference on Computer Vision and Pattern Recognition (CVPR)* (Institute of Electrical and Electronics Engineers, 2016), pp. 770–778.
65. A. Jagadeesh, J. L. Gardner, Texture-like representations of objects in human visual cortex. Open Science Framework. https://osf.io/gpx7y/. Deposited 3 April 2022.
66. K. J. Friston *et al.*, Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.* **2**, 189–210 (1994).
67. K. N. Kay, A. Rokem, J. Winawer, R. F. Dougherty, B. A. Wandell, GLMdenoise: A fast, automated technique for denoising task-based fMRI data. *Front. Neurosci.* **7**, 247 (2013).

68. S. O. Dumoulin, B. A. Wandell, Population receptive field estimates in human visual cortex. *Neuroimage* **39**, 647–660 (2008).
69. A. Stigliani, K. S. Weiner, K. Grill-Spector, Temporal processing capacity in high-level visual cortex is domain specific. *J. Neurosci.* **35**, 12412–12424 (2015).
70. M. F. Glasser *et al.*, A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
71. N. J. Majaj, H. Hong, E. A. Solomon, J. J. DiCarlo, Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *J. Neurosci.* **35**, 13402–13418 (2015).
72. M. Schrimpf *et al.*, Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* **108**, 413–423 (2020).
73. L. T. Maloney, J. N. Yang, Maximum likelihood difference scaling. *J. Vis.* **3**, 573–585 (2003).
74. M. N. Hebart, C. Y. Zheng, F. Pereira, C. I. Baker, Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nat. Hum. Behav.* **4**, 1173–1185 (2020).
75. T. Bonnen, D. L. K. Yamins, A. D. Wagner, When the ventral visual stream is not enough: A deep learning account of medial temporal lobe involvement in perception. *Neuron* **109**, 2755–2766.e6 (2021).
76. I. Ustyuzhaninov, W. Brendel, L. A. Gatys, M. Bethge, Texture synthesis using shallow convolutional networks with random filters. arXiv [Preprint] (2016). https://doi.org/10.48550/arXiv.1606.00021 (Accessed 7 April 2022).
77. P. Thompson, Margaret Thatcher: A new illusion. *Perception* **9**, 483–484 (1980).
78. J. W. Tanaka, M. J. Farah, Parts and wholes in face recognition. *Q. J. Exp. Psychol. A* **46**, 225–245 (1993).
79. M. J. Farah, K. D. Wilson, M. Drain, J. N. Tanaka, What is "special" about face perception? *Psychol. Rev.* **105**, 482–498 (1998).
80. W. Brendel, M. Bethge, Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. arXiv [Preprint] (2019). https://doi.org/10.48550/arXiv.1904.00760 (Accessed 7 April 2022).
81. G. K. Aguirre, E. Zarahn, M. D'Esposito, An area within human ventral cortex sensitive to "building" stimuli: Evidence and implications. *Neuron* **21**, 373–383 (1998).
82. R. Epstein, N. Kanwisher, A cortical representation of the local visual environment. *Nature* **392**, 598–601 (1998).
83. S. Nasr, C. E. Echavarria, R. B. H. Tootell, Thinking outside the box: Rectilinear shapes selectively activate scene-selective cortex. *J. Neurosci.* **34**, 6721–6735 (2014).
84. M. Huh, P. Agrawal, A. A. Efros, What makes ImageNet good for transfer learning? arXiv [Preprint] (2016). https://doi.org/10.48550/arXiv.1608.08614 (Accessed 7 April 2022).
85. S. Kornblith, J. Shlens, Q. V. Le, "Do better ImageNet models transfer better?" in 2019 *Conference on Computer Vision and Pattern Recognition (CVPR)* (Institute of Electrical and Electronics Engineers, 2019), pp. 2656–2666.
86. N. Kriegeskorte, M. Mur, P. Bandettini, Representational similarity analysis–Connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
87. D. L. K. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
88. B. A. Richards *et al.*, A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).
89. S. Ogawa *et al.*, Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 5951–5955 (1992).
90. B. A. Wandell, J. Winawer, Imaging retinotopic maps in the human brain. *Vision Res.* **51**, 718–737 (2011).
91. J. L. Gardner, E. P. Merriam, J. A. Movshon, D. J. Heeger, Maps of visual space in human occipital cortex are retinotopic, not spatiotopic. *J. Neurosci.* **28**, 3988–3999 (2008).
92. J. Larsson, D. J. Heeger, Two retinotopic visual areas in human lateral occipital cortex. *J. Neurosci.* **26**, 13128–13142 (2006).