Mid-level visual features underlie the high-level categorical organization of the ventral stream

Bria Long[a,b,1], Chen-Ping Yu[a,c], and Talia Konkle[a]

[a]Department of Psychology, Harvard University, Cambridge, MA 02138; [b]Department of Psychology, Stanford University, Stanford, CA 94305; and [c]Phiar Technologies, Inc., Palo Alto, CA 94303

Edited by Tony Movshon, New York University, New York, NY, and approved August 7, 2018 (received for review November 10, 2017)

Human object-selective cortex shows a large-scale organization characterized by the high-level properties of both animacy and object size. To what extent are these neural responses explained by primitive perceptual features that distinguish animals from objects and big objects from small objects? To address this question, we used a texture synthesis algorithm to create a class of stimuli—texforms—which preserve some mid-level texture and form information from objects while rendering them unrecognizable. We found that unrecognizable texforms were sufficient to elicit the large-scale organizations of object-selective cortex along the entire ventral pathway. Further, the structure in the neural patterns elicited by texforms was well predicted by curvature features and by intermediate layers of a deep convolutional neural network, supporting the mid-level nature of the representations. These results provide clear evidence that a substantial portion of ventral stream organization can be accounted for by coarse texture and form information without requiring explicit recognition of intact objects.

ventral stream organization | mid-level features | object recognition | fMRI | deep neural networks

The ventral visual stream transforms retinal input into representations that help us recognize the categories of objects in the visual world (1, 2). The structure of this cortex has been characterized at various levels of granularity. For a few specific categories—faces, bodies, scenes, and visual words—there is a mosaic of highly selective neural regions in occipitotemporal cortex (3–6). Other basic-level category distinctions (e.g., shoes vs. keys) lack clear category-specific regions, but they also can be decoded from multivoxel patterns in this same cortex (7, 8). Even broader categorical distinctions reflecting the animacy and real-world size of objects are evident in the large-scale spatial structure of occipitotemporal cortex (9–13). While these organizing dimensions of the ventral stream are well documented, understanding the nature of the visual feature tuning underlying these ubiquitous categorical responses and their spatial organization across the cortex has proven notoriously difficult.

One key challenge is methodological: Any measured neural response to recognizable object categories may actually reflect the processing of low-level image statistics, mid-level perceptual features, holistic category features, or even semantic associations that are not visual at all (or some combination of these features). In other words, there is a continuum of possible representational levels that could account for neural responses to object categories. Within a classic view of the ventral visual hierarchy (1, 14) there is broad agreement that low-level features are processed in early visual regions and high-level, categorical inferences take place in later, downstream regions, including the anterior temporal lobe (15). However, for the neural representations in intermediate occipitotemporal cortex, there is active debate about just how "high" or "low" the nature of the representation is.

At one extreme, some evidence suggests that the categorical neural responses are quite high-level, reflecting the interpretation of objects as belonging to a given category rather than anything about their visual appearance per se (see ref. 16 for a recent review). For example, when ambiguous moving shapes are identified as "animate," they activate a cortical region that prefers animals (17, 18). Within the inanimate domain, hands-on training to treat novel objects as tools increases neural responses to these novel objects in tool-selective areas (19). In addition, differences between object categories persist when attempting to make them look as similar as possible [e.g., a snake vs. a rope (20–22), but see ref. 23 for critiques to this approach]. These findings and others from congenitally blind participants (24–28) have led to the strong claim that visual features are insufficient to account for categorical responses in visual cortex (16).

At the same time, a growing body of work demonstrates that neural responses in occipitotemporal cortex also reflect very low-level visual information. Retinotopic maps are now known to extend throughout high-level visual cortex (29–34). Furthermore, low-level visual features such as luminance and the presence of rectilinear edges account for a surprising amount of variance in neural responses to objects (35, 36). More recently, some evidence suggests that recognizable objects and unrecognizable, locally phase-scrambled versions of objects yield similar neural patterns across occipitotemporal cortex (37, but see ref. 38). Taken together, these results have led to an alternative proposal in which the categorical responses of occipitotemporal cortex are solely a byproduct of simple low-level visual-feature maps and are not related to the categories per se (39, 40).

These two current viewpoints represent two prominent models of how to characterize the representations in occipitotemporal cortex. In an intermediate account, neural responses in occipitotemporal cortex reflect tuning to visual features of intermediate complexity (e.g., refs. 13 and 41–43). That is, it is mid-level features, combinations of which reflect the "shape of things" (7), that underlie categorical responses. However, neural evidence for a mid-level feature representation is sparse, in part because there is no widely accepted model of mid-level features. For example, is the basis set of this

Significance

While neural responses to object categories are remarkably systematic across human visual cortex, the nature of these responses has been hotly debated for the past 20 y. In this paper, a class of stimuli (texforms) is used to examine how mid-level features contribute to the large-scale organization of the ventral visual stream. Despite their relatively primitive visual appearance, these unrecognizable texforms elicited the entire large-scale organizations of the ventral stream by animacy and object size. This work demonstrates that much of ventral stream organization can be explained by relatively primitive mid-level features without requiring explicit recognition of the objects themselves.

Author contributions: B.L. and T.K. designed research; B.L. and C.-P.Y. performed research; B.L., C.-P.Y., and T.K. contributed analytic tools; B.L., C.-P.Y., and T.K. analyzed data; and B.L. and T.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: All preprocessed fMRI data are available at https://osf.io/69pbd/.

[1]To whom correspondence should be addressed. Email: brialorelle@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1719616115/-/DCSupplemental.

feature space derived from generic building blocks (i.e., ref. 44) or from features tightly linked to categorical distinctions (e.g., the presence of eyes)? As such, isolating mid-level representations and mapping their relationship to categorical responses are challenging both methodologically and theoretically.

Here, we approached this challenge by leveraging a class of stimuli—"texforms." Specifically, we used a texture-synthesis algorithm to generate synthetic stimuli which capture some texture and coarse form information from the original images but to most people look like "texturey blobs" (Fig. 1 and *SI Appendix*, Fig. S1) (45–48). These stimuli have two properties that make them particularly well suited for probing neural levels of representation along the visual hierarchy. First, people cannot identify what these stimuli are at the basic level (e.g., as a "cat"); thus texforms clearly lack some critical high-level features, i.e., those that enable basic-level categorization (*SI Appendix*, Fig. S2). However, even though texforms are not identifiable, they do retain some statistical visual information related to the broad classes of animals vs. objects and big objects vs. small objects, distinctions that are known to structure the large-scale organization of occipitotemporal cortex (10, 11). For example, participants seem to rely in part on the perceived curvature of a texform to guess above chance whether it is animate or inanimate and whether it is big or small in the world (*SI Appendix*, Figs. S3 and S4) (46–48). Thus, with this stimulus set we are now poised to ask whether the features preserved in these texform stimuli are sufficient to drive neural differences between animals and objects of different sizes and where along the ventral stream any differences manifest.

To anticipate our results, we find clear evidence that the mid-level perceptual features preserved in texforms are sufficient to drive the ventral stream organization by animacy and object size. Surprisingly, these differences manifested extensively throughout the entire occipitotemporal cortex, driving even more anterior, purportedly "high-level" regions. To better understand the nature of the visual representation in this cortex, we used a model comparison approach, testing how well a variety of image-feature models could predict the structure in the neural responses to both texforms and recognizable images. These analyses revealed that both perceived curvature ratings and the intermediate visual features learned by deep convolutional neural networks (CNNs) (49) were able to explain a substantial portion of the variance in neural response patterns to texforms and recognizable images; in contrast, models based on low-level image statistics fit poorly. These results demonstrate that animacy and object size responses in occipitotemporal cortex can be explained to a large degree by mid-level perceptual features including texture and coarse form information. We propose that mid-level features meaningfully covary with high-level distinctions between object categories and that this relationship underlies the large-scale organization of the ventral stream.

Results

Observers viewed texform images of big objects, small objects, big animals, and small animals, followed by their recognizable counterpart images, while undergoing functional neuroimaging. Fig. 1 shows the full stimulus set. All images were presented in a standard blocked design, enabling us to examine the univariate effects of the two main dimensions (animacy and size) for both texforms and original image sets.

Additionally, we included a nested factor in the design related to texform classifiability. Specifically, for each texform, a classifiability score was calculated based on how well a separate set of observers could guess its animacy and real-world size (*SI Appendix*, Fig. S3). These scores were used to vary the classifiability of each block of texforms systematically (*Methods*), and original images were also presented in the same groups in yoked runs. This nested factor created a secondary, condition-rich design, enabling us to examine the structure of multivoxel patterns to texforms and original images. Importantly, subjects in the neuroimaging experiment were never asked to identify or classify the texforms and were not even informed that they were viewing pictures generated from recognizable images (see also *SI Appendix*, Fig. S5).

Animacy and Object Size Topographies. To examine whether texforms elicited an animacy organization, we compared all animal and object texform univariate responses in each participant by plotting the difference in activation within a visually active cortex mask (all > rest, $t > 2$) (*SI Appendix*, Fig. S6). Systematic differences in response to animal versus object texforms were observed across the entire occipitotemporal cortex, with a large-scale organization in their spatial distribution. The

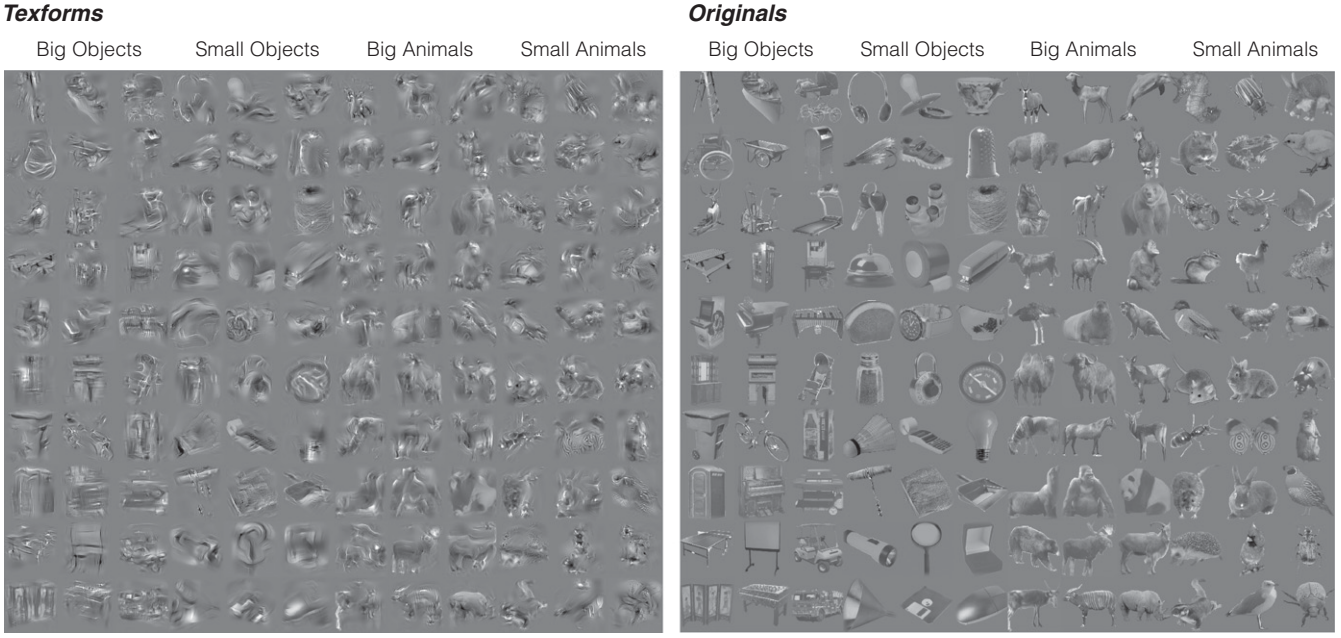

Fig. 1. Texforms (*Left*) were generated using a texture-synthesis model (45) from recognizable pictures (*Right*) of 30 big objects, 30 small objects, 30 big animals, and 30 small animals. Stimuli are shown at slightly higher contrast for visualization purposes. Stimuli were selected so that all texforms were unrecognizable at the basic level using online recognition experiments.

same analysis was conducted using responses measured when observers viewed the original, recognizable images of animals and objects. The preference maps for both texforms and original animacy organizations are shown for a single subject in Fig. 2*A* and reveal a striking degree of similarity (see group topographies in *SI Appendix*, Fig. S7 and all single-subject topographies in *SI Appendix*, Fig. S8). Thus, even though there is an obvious perceptual difference between texforms and recognizable objects, they elicit similar topographies along the entire occipitotemporal cortex.

To quantify this correspondence, we computed the correlation between the original and texform preference maps separately in each participant within active occipitotemporal voxels following ref. 50 (*Methods*). The map correlation coefficients for each participant and the average correlation coefficient for the group are plotted in Fig. 2*B*. Overall, voxels in occipitotemporal cortex had similar animacy preferences for recognizable and texform images in all subjects, resulting in a robust correlation at the group level (average $r = 0.74$, SD = 0.07, permutation test significant in each subject, all $P < 0.001$; average noise ceiling across subjects, $r = 0.81$, SD = 0.06) (*Methods*).

Next, we used a similar analysis to examine whether texforms also elicited a real-world size organization. Given that the size organization is found only for inanimate objects, not animals (10), we compared the responses to big objects versus small objects. Note that this yields half the power in the design to examine the object size organization relative to the animacy organization. Nonetheless, big and small object texforms elicited robust differential responses along the ventral stream, with a systematic large-scale spatial organization similar to that elicited by original images (Fig. 2*C*; see group topographies in *SI Appendix*, Fig. S7 and all single-subject topographies in *SI Appendix*, Fig. S8). Quantitatively, moderate correlations between original and texform preference maps were found in all but one participant, resulting in robust map correlations at the group level (average $r = 0.41$, SD = 0.20, permutation test significant in seven of eight subjects at $P < 0.001$) (*Methods* and Fig. 2*D*). While the overall magnitude of the object size group map correlation was weaker than the animacy map correlation, note that the noise ceiling of the data was lower, likely reflecting the fact that half the data were used in this analysis (average noise ceiling across subjects, $r = 0.39$, SD = 0.32) (Fig. 2*D*).

Given that some texforms are better classified by their animacy and real-world size, do these better-classified texforms elicit spatial topographies that are even more similar to those for the original images? To examine this possibility, we split the data in half by texform classifiability. For the animacy distinction, map correlations between original images and texforms were higher for better-classified texforms (average $r = 0.73$) than for more poorly classified texforms [$M = 0.45$, $t(7) = 7.00$, $P < 0.001$]. However, the size organization was not as strongly influenced by classifiability [average map correlation for better-classified texforms vs. original images: $M = 0.35$; poorly classified texforms vs. original images: $M = 0.29$; $t(7) = 1.25$, $P = 0.25$].

There are at least two possible reasons for this result. On one hand, better-classified texforms could drive stronger animacy responses because neural responses to better-classified texforms are amplified by top-down feedback from other regions that process semantic information. However, an alternative possibility is that better-classified texforms also better preserve the relevant textural and curvature statistics of animals and objects (47, 48). We return to this effect of texform classifiability on neural responses in the predictive modeling section, exploring in detail why more classifiable texforms might drive differential neural responses.

Posterior-to-Anterior Analysis. Within a classic view of the ventral stream hierarchy, posterior representations reflect more primitive features, and anterior representations reflect more sophisticated features. We next looked for evidence of this hierarchy with respect to the animacy and object-size organizations, specifically examining whether original images evoked stronger animacy and size preferences than texforms in more anterior regions. To do so, we defined five increasingly anterior regions of occipitotemporal cortex using anatomical coordinates (*Methods* and Fig. 3).

We first looked at animacy preferences along this gradient to determine whether the category preference becomes increasingly larger for original images (vs. texforms) in more anterior regions.

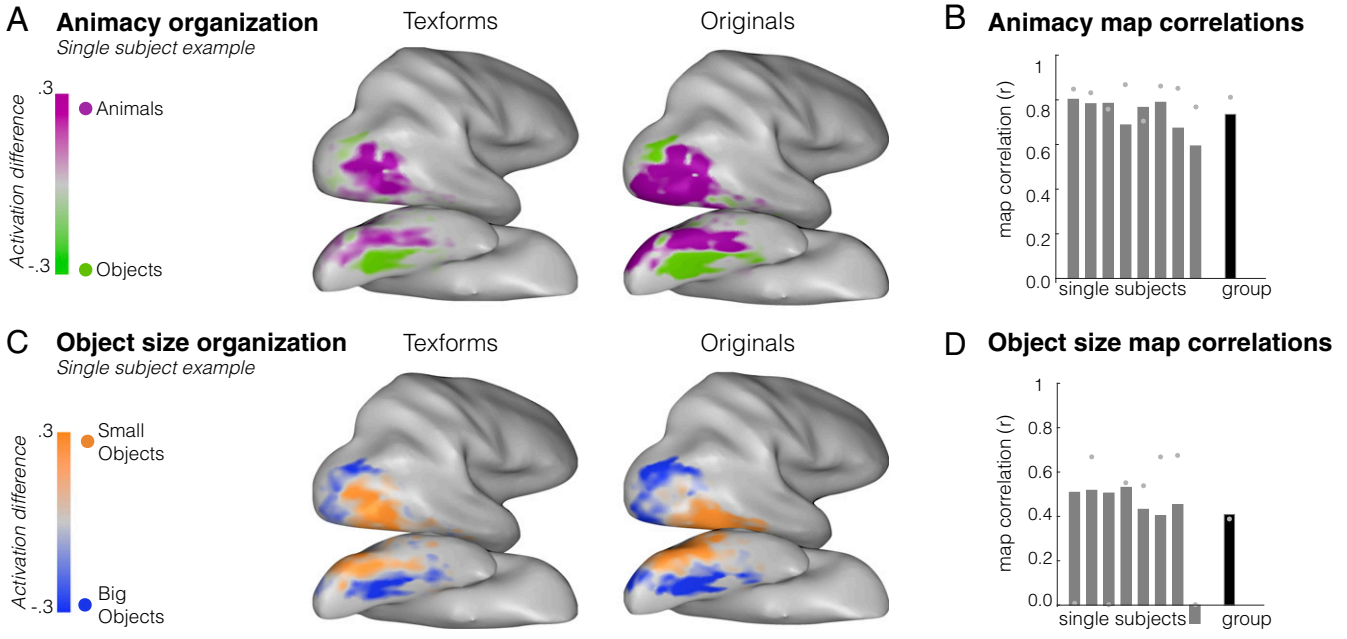

Fig. 2. Preference map analyses. (*A* and *C*) Response preferences within active occipitotemporal voxels are plotted for animals vs. objects (*A*) and for big vs. small objects (*C*) in an example participant, considering texform images (*Left*) and original images (*Right*). The color bar reaches full saturation at activation differences between 0.3 and −0.3 (reflecting the beta difference calculated from this individual's GLM). (*B* and *D*) The correlation between the original and texform response maps in active occipitotemporal voxels is plotted for the animacy (*B*) and object size (*D*) distinctions. Correlations between original image and texform image maps are shown for all individual participants and for the group, averaged across all subjects. Gray dots indicate the estimated noise ceiling for each participant and at the group level.

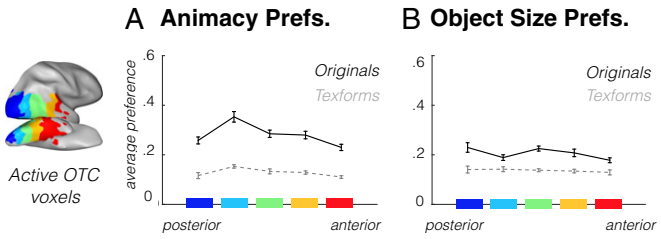

Fig. 3. Anatomical sections (shown here at the group level) from posterior to anterior in blue to red. (*A*) The strength of the average animal/object response preference is shown for each anatomical section, averaged across voxels and participants, plotted for both original images (black solid line) and texform images (gray dashed line). Error bars reflect between-subjects SEM. (*B*) The strength of the average big/small object response preference is shown, as in *A*.

The overall strength of the animal/object preferences in each sector is shown in Fig. 3*A* (see also *SI Appendix*, Fig. S9), where the solid lines show the average preference strength for original images, and the dashed lines show the average preference strength for texform images. If original images exhibited stronger category preferences in more anterior regions, this would be evident by an increasing difference between the solid and dashed lines. Instead, these two lines are relatively parallel [average activation difference for each section in animacy preferences for original images vs. texforms: $M_{S1} = 0.14$, $M_{S2} = 0.20$, $M_{S3} = 0.15$, $M_{S4} = 0.15$, $M_{S5} = 0.12$; average rank correlation across subjects between these activation differences and anatomical sections: $r = -0.34$, t test against zero, $t(7) = -1.92$, $P = 0.10$] (Fig. 3*A*). Thus, this analysis reveals that original images generate stronger category preferences than do texforms across all anatomical sections, not only in more anterior ones.

When we conducted the same analyses on the object size distinction, we found the same pattern of effects. That is, original images elicited stronger big/small object preferences than texforms across all anatomical sections, and this difference was relatively consistent from posterior to anterior sections [average activation difference for each section in object size preferences for original images vs. texforms: $M_{S1} = 0.09$, $M_{S2} = 0.05$, $M_{S3} = 0.09$, $M_{S4} = 0.07$, $M_{S5} = 0.05$; average rank correlation between these activation differences and anatomical sections: $r = -0.25$, t test against zero, $t(7) = -1.17$, $P = 0.28$] (Fig. 3*B*). In other words, we found little if any evidence for the pattern of results that might be expected from a simple visual hierarchy in which texforms and original neural responses matched in posterior areas but diverged in anterior areas. Instead, the difference in animacy/size preferences for original images vs. texforms remained relatively constant across the full posterior-to-anterior gradient.

One possible factor that might influence the interpretation of this result is the overall neural activity: Perhaps original images simply drive all voxels along the ventral stream more than texforms, and thus the greater animacy/size preferences we observe for original images actually reflect a better signal-to-noise ratio. If so, the original and texform organizations may be even more similar to each other than we have measured. To examine this possibility, we analyzed the overall magnitude of the neural response to original images vs. texforms along this posterior-to-anterior axis, averaging across all animacy/size conditions. Overall, voxels were driven relatively similarly by both texforms and original images across all anatomical sections, although recognizable images generated slightly more overall activity than texforms in the more anterior sections [average activation difference between original images and texforms in each section: $M_{S1} = -0.01$, $M_{S2} = 0.07$, $M_{S3} = 0.12$, $M_{S4} = 0.11$, $M_{S5} = 0.14$; average rank correlation between activation differences and anatomical sections: $r = 0.59$, t test against zero, $t(7) = 3.84$, $P = 0.006$] (*SI Appendix*, Fig. S10). Thus, it was not the case that original images elicited stronger overall responses everywhere, and the response magnitude is unlikely to explain away the result that original images elicit stronger animacy/object and big/small object preferences across the ventral stream.

In sum, both texforms and recognizable images generated large-scale topographies by animacy and object size throughout the entire ventral stream, with recognizable images generating overall stronger category preferences (*SI Appendix*, Fig. S11). We did not find strong evidence for a hierarchy of representations that differentiated between texforms and recognizable images. Instead, these results point toward mid-level features as a major explanatory factor for the spatial topography of object responses along the entire occipitotemporal cortex.

Tolerance of Retinal Position. Given the extensive activation of these texforms along the ventral stream, one potential concern is that these texform topographies may reflect simple retinotopic biases that also extend throughout this cortex rather than mid-level feature information per se. For example, if animal texforms happen to have more vertical information in the lower visual field, and object texforms have more horizontal information in the upper visual field, then such low-level retinotopic differences might account for the responses observed in the main experiment. To test this possibility, we conducted a second experiment in which a different group of observers was shown the same stimuli (both texforms and recognizable images), but each image was presented separately above and below fixation (*SI Appendix*, Fig. S12). If animacy and size preferences are maintained over changes in visual field position, this provides evidence against a simple retinotopic explanation.

In our first analysis, we examined how much of occipitotemporal cortex showed location-tolerant animacy and size preferences separately for original images and texforms. To do so, animal vs. object preferences were computed separately when images were presented in the upper visual field location and in the lower visual field location. We retained voxels that showed the same category preference (e.g., animals > objects or objects > animals) when stimuli were presented in the upper visual field and when stimuli were presented in the lower visual field (*SI Appendix*, Fig. S13). The percent of retained voxels relative to the total set of active occipitotemporal cortex voxels was computed for both the animacy and object size distinctions for both original and texform images separately in each participant.

When subjects viewed the original images, we found that 71% (SD = 4%) of voxels in occipitotemporal cortex showed location-tolerant animacy preferences, and 55% (SD = 8%) of voxels showed location-tolerant object size preferences. When subjects viewed texforms, we found that 56% (SD = 13%) of occipitotemporal voxels showed location-tolerant animacy preferences, and 47% (SD = 5%) of voxels showed location-tolerant object size preferences. Thus, both recognizable images and texforms elicited animacy and object size preferences that were largely tolerant to changes in visual field position.

Next, we assessed the similarity of category preferences elicited by texform and original images within these location-tolerant voxels. That is, do the voxels that show location-tolerant preferences for animacy and size when subjects view original images show the same category preferences when subjects view texforms? Animacy/object size topographies for texforms/original images are shown within these location-tolerant conjunction voxels in Fig. 4*A* and qualitatively show similar spatial profiles (see group topographies in *SI Appendix*, Fig. S14 and all single-subject topographies in *SI Appendix*, Fig. S15). Quantitatively, we again conducted map correlations within voxels that showed consistent category preferences across retinal locations for original images. Texform and original topographies again showed a high degree of spatial correspondence within these location-tolerant voxels, evident in single subjects and at the group level (animacy: average $r = 0.67$, SD = 0.12; size: average $r = 31$, SD = 0.11, permutation tests against shuffled voxel baseline significant in all subjects at $P < 0.001$) (Fig. 4*B*). Furthermore, when we relaxed our voxel-inclusion criterion, analyzing map correlations within all visually active voxels in occipitotemporal cortex, as in experiment 1, we found

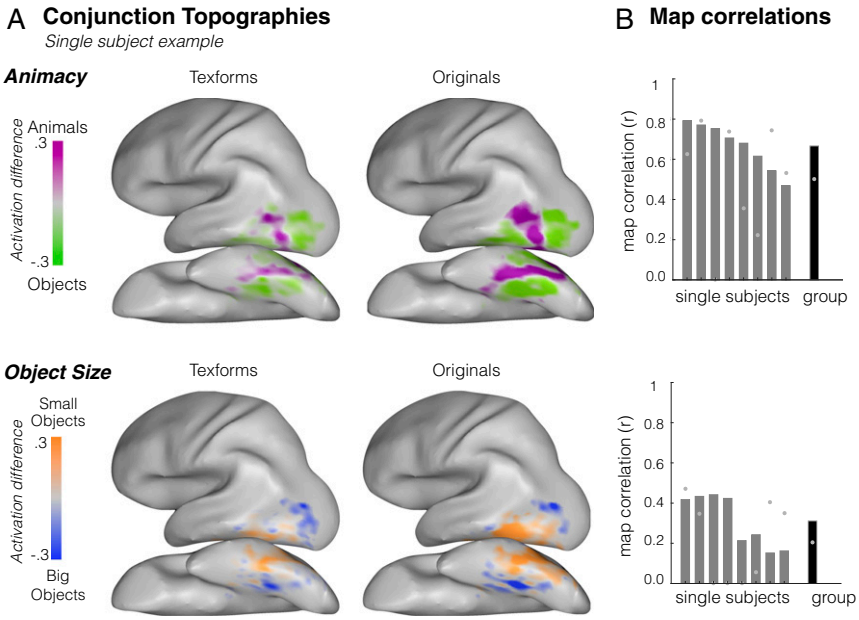

Fig. 4. (*A*). Group conjunction topographies. Average category responses when texforms (*Left*) and original images (*Right*) are presented in the upper and lower visual field. Topographies are restricted to voxels that show the same category preference regardless of the stimuli's location in the visual field and are shown separately for animacy (*Upper*) and size (*Lower*). (*B*) Conjunction map correlation values (*y* axis) are plotted for each individual subject (*x* axis) and at the group level separately for animacy (*Upper*) and object size (*Lower*) contrasts; gray dots indicate the noise ceiling for each participant and at the group level.

the same pattern of results (animacy: average $r = 0.60$, SD = 0.11; size: average $r = .25$, SD = 0.11), indicating that the stringent voxel-inclusion criterion did not bias the results.

Compared with the initial experiment, the organizations found in the second experiment are sparser, particularly for the object size distinction. This may indicate stronger retinotopic contributions for the object size relative to the animacy distinction or may simply reflect the lower signal-to-noise ratio in the object size analysis (for which only half the data are used). Nonetheless, these results demonstrate that these topographies reflect mid-level information that is tolerant of changes in visual field position, replicating and extending the primary finding.

Predictive Modeling: Texforms. We next aimed to provide insight into the nature of the mid-level features that actually drive these animacy and size texform response differences across the ventral stream. To do so, we compared how well a variety of models predict the multivoxel pattern similarity to groups of texforms across occipitotemporal cortex (51–53).

We first constructed representational dissimilarity matrices (RDMs) in occipitotemporal cortex using data from the richer condition structure nested in our experiment design (*SI Appendix*, Fig. S16). Recall that every time observers saw a block of texform images, this block was comprised of a set of texforms from one of six levels of classifiability. The more classifiable the texform, the better a separate group of norming participants was able to guess that this texform was generated from an animal versus an object, or from a big versus small thing in the world (*SI Appendix*, Fig. S3). Examples of well-classified and poorly classified texforms (and their accompanying original counterparts) are shown in Fig. 5*A*.

Fig. 5*B* shows the similarity in the multivoxel patterns elicited by texforms and the corresponding original images. The texform RDM has some gradation in texform levels of classifiability, which by inspection shows that more classifiable texforms are more dissimilar from each other. By comparison, the structure in responses to recognizable images is more categorical in nature, with a clear animate/inanimate division that is visually evident in the quadrant structure of the RDM and with a weaker but visible big/small object division in the upper left quadrant.

What features best predict this neural similarity structure generated by texforms? Here, we tested the predictive power of a range of feature spaces, including basic image statistics, activations in each layer of a deep CNN (49), and behavioral ratings of perceived curvature, using a weighted representational modeling procedure as introduced in ref. 51. This procedure entailed constructing RDMs for each feature in a given feature space and weighting the individual features to best predict the group neural RDM. Model performance was cross-validated using an iterative procedure (*Methods* and ref. 51). The key outcome measure is the degree to which this predicted neural RDM matches the observed neural data in each subject (using rank correlation Kendall tau-a, τ_A). All model performance is put in the context of the neural noise ceiling, reflecting how well a given subject's RDM can predict the group RDM (*Methods*, Fig. 6*A*, and ref. 52).

Basic Image Statistics. First, we examined how well combinations of low-level image statistics could account for the observed neural structure. While some prior work has found such statistics to be an

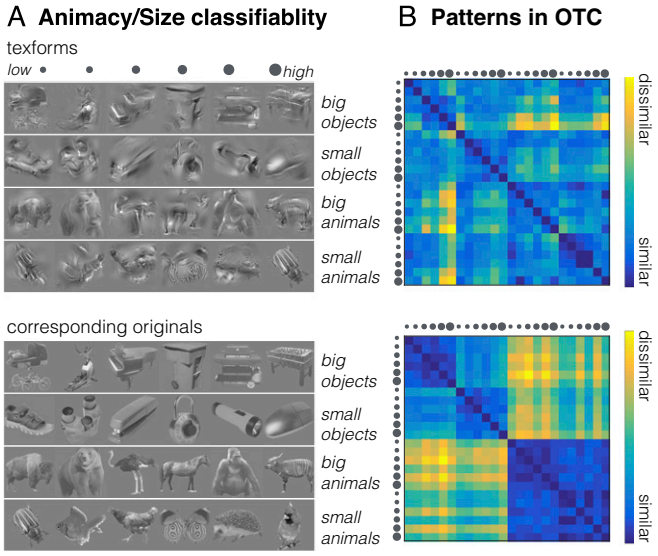

Fig. 5. (*A*, *Upper*) Examples of texforms from the six classifiability groups, from lowest to highest, are shown for the four main conditions. (*Lower*) The corresponding original images. (*B*) Representational dissimilarity matrices obtained from neural patterns in the active occipitotemporal cortex for texforms (*Upper*) and original images (*Lower*). Data are scaled so that in both cases the most dissimilar values are yellow and the least dissimilar values are blue.

insufficient basis for predicting the geometrical layout of categorical responses in the inferior temporal cortex (53, 54), others have argued for their sufficiency (37, 55). With our cross-validation modeling procedure, we found that weighted linear combinations of the texture-synthesis model features predicted relatively little variance in the neural patterns relative to the noise ceiling [$\tau_A = 0.16$; noise ceiling $\tau_A = (0.38–0.48)$]. Consistent with this result, other models based on low-level image statistics also captured only a small amount of variance (Gabor model, $\tau_A = 0.12$; Gist model, $\tau_A = 0.10$) (56). Thus, linear combinations of relatively simple visual features were not sufficient to predict the multivoxel patterns of occipitotemporal cortex.

Convolutional Neural Network. We next tested models constructed from deep CNN unit responses, reflecting the state of the art in predicting neural responses to objects (53, 57). To do so, we extracted the feature representations throughout all layers of a CNN (AlexNet) (*SI Appendix*) in response to texforms. Note that while this CNN was pretrained to categorize 1,000 object categories, it was not specifically trained on (or tuned to) any of the texform images or their recognizable counterparts.

Models constructed from representations in the earliest layers of a CNN performed poorly, similar to the models based solely on image statistics. However, predictive ability increased through the first few convolutional layers, plateauing around convolutional layers 4 and 5 (layer 4: $\tau_A = 0.31$; layer 5: $\tau_A = 0.32$). Thus, the variation in neural patterns in response to different groups of texforms was relatively well captured by responses in mid-level convolutional layers of a deep CNN. These results reveal that mid-level features captured by these intermediate CNN layers can explain the variation in neural patterns in response to different groups of texforms.

Curvature Ratings. We next asked how well the perceived curvature ratings could explain this neural structure, based on behavioral evidence that boxy/curvy ratings distinguish animals, small objects, and big objects (*SI Appendix*, Fig. S4) (46–48) and in line with a growing body of work implicating curvature as a critical mid-level feature in ventral stream responses (36, 50, 58). We found that this simple, one-dimensional model based on curvy/boxy judgments was able to predict the structure moderately well ($\tau_A = 0.28$), capturing almost 50% of the variance in the neural patterns elicited by texforms.

Animacy/Size Classification. As a sanity check, we examined the performance of a behavioral model constructed directly from the classification scores used to group the texforms into the nested conditions by classifiability. We expected this model to perform well, as we built this structure into our experiment design. Overall, we found that these animacy/size judgments were able to predict the structure of texform responses near the noise ceiling [average subject RDM-to-model correlation, $\tau_A = 0.38$; noise ceiling $\tau_A = (0.38–0.48)$]. This result confirms that the neural patterns in response to texforms varied as a function of the classifiability of the texforms; groups of texforms that were better classified by their animacy/size elicited more distinct neural patterns.

A summary of these texform modeling results in occipitotemporal cortex is shown in Fig. 6*A*. To visually inspect the structure captured by the different models, predicted neural RDMs from several models are shown. The overall performance for all models, reflecting the average model-to-subject RDM correlation, is shown in the bar plot. Taken together, these analyses show that models based on intuitive curvature ratings and intermediate layers of a deep CNN captured this neural structure relatively well, while models based on early CNN layers and simple image statistics were insufficient. Broadly, these modeling results provide computational support for the mid-level nature of this neural representation and help triangulate the kinds of features that drive neural responses to texforms in occipitotemporal cortex (i.e., curvy/boxy mid-level features of intermediate complexity).

The success of the CNN modeling also helps clarify the role that texform classifiability has on neural responses. Specifically, one potential factor in interpreting neural responses to texforms is that the more classifiable texforms may engender feedback such that top-down effects could contribute to the apparent organization (e.g., some evidence for an animal causes attentional amplification of animal-related regions). However, CNN responses to texforms were able to predict neural responses to texforms relatively close to the noise ceiling. Critically, the CNN does not have top-down feedback and thus has no mechanism by which to amplify any animacy/size differences (see also *SI Appendix*, Figs. S17 and S18 for

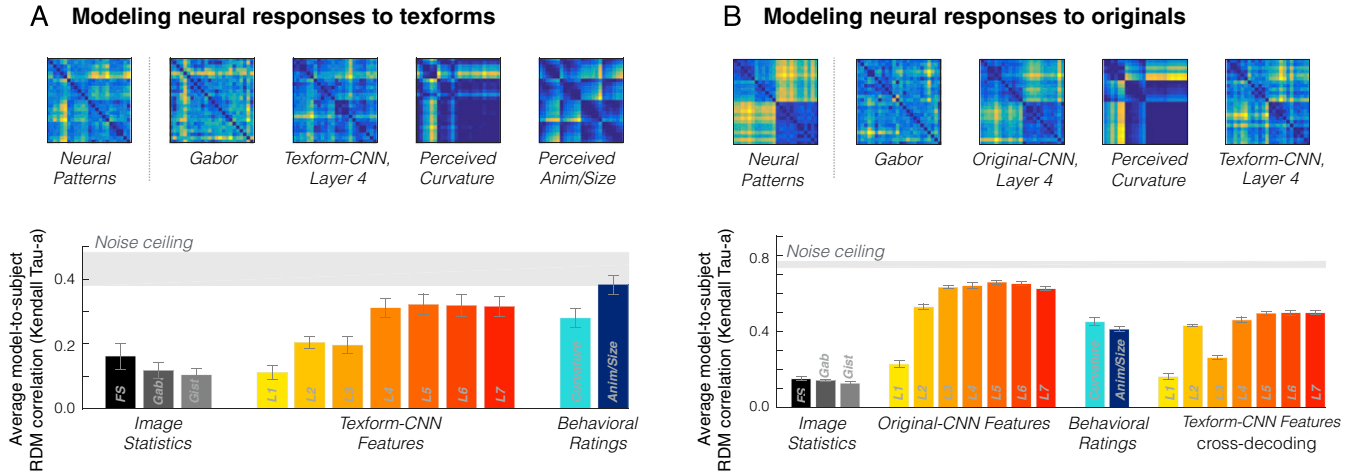

Fig. 6. (*A*, *Upper*) Neural patterns in response to texforms (shown in Fig. 5*B*) and predicted neural dissimilarities for selected models obtained through the cross-validation procedure. (*Lower*) The bar plot shows the predicted model correlation (Kendall τ_A). Error bars reflect the SE of the model fit across individual subject's neural patterns in occipitotemporal cortex. The bars show different models, from left to right: Freeman and Simoncelli texture model (black), Gabor model (dark gray), Gist model (light gray), AlexNet features layers 1–7 (yellow to red), curvature behavioral ratings (light blue), and animacy/size behavioral ratings (dark blue). Data are plotted with respect to the noise ceiling of neural responses to texform images across participants, shown in light gray. (*B*, *Upper*) Neural patterns in response to original images (shown in Fig. 5*B*) and predicted neural dissimilarities for four models obtained through the same leave-one-condition-out cross-validation procedure. (*Lower*) The average predicted model correlation (Kendall τ_A) is plotted for different models, as in *A*, with AlexNet features extracted from both original images and texforms. Data are plotted with respect to the noise ceiling of neural responses to original images across participants, shown in light gray.

modeling results in each anatomical section of occipitotemporal cortex). Thus, the modeling results are consistent with the idea that some texforms are more classifiable than others because they preserve more of the relevant mid-level features.

Predictive Modeling: Recognizable Images and Cross-Decoding. For completeness, we also compared how well the same set of models could predict the structure of neural responses to original images; the results are summarized in Fig. 6*B* (see also *SI Appendix*, Fig. S19). Overall, we found a pattern of results similar to that we found with texforms: Basic image statistic models performed poorly [Freeman & Simoncelli features: $\tau_A = 0.15$; Gabor features: $\tau_A = 0.14$; Gist model: $\tau_A = 0.13$; neural noise ceiling $\tau_A = (0.73–0.77)$], while the feature representations elicited in deep CNNs by recognizable images almost fully predicted these neural patterns by intermediate layers (layer 4: $\tau_A = 0.64$; layer 5: $\tau_A = 0.66$). Interestingly, as with texforms, a model based on curvy/boxy judgments of the original images also accounted for a substantial portion of the variance ($\tau_A = 0.45$). Finally, a categorical model of animacy and object size only performed moderately well ($\tau_A = 0.41$), consistent with prior work highlighting that occipitotemporal cortex has a more graded similarity structure (53). Taken together, the predictive power of the intermediate CNN layers and the curvature ratings suggest that mid-level representation underlies a substantial component of neural responses to recognizable images.

We next performed a stronger test of this argument by conducting a cross-decoding analysis. Specifically, we examined whether neural responses to original images could be predicted using the CNN features extracted from texforms. In other words, we tested whether the neural similarity of original images could be predicted from deep neural network responses to the texform counterparts of each original image. Indeed, CNN texform features were able to predict much of the RDM structure elicited by recognizable images (layer 4: $T_A = 0.46$) (Fig. 6*B*). This cross-decoding analysis further supports the idea that the neural responses to recognizable objects are driven substantially by mid-level feature information.

Discussion

We employed a stimulus set—texforms—to examine if and how mid-level features contribute to the large-scale organization of the ventral stream. We found that (*i*) texform stimuli were sufficient to elicit animacy and size topographies along occipitotemporal cortex, well into what are classically considered higher-level object-selective areas; (*ii*) these mid-level topographies were not inherited from low-level retinotopic biases, as they generalized over the visual field position; (*iii*) the similarity in the structure of the neural representations of both texforms and recognizable images was best predicted by intermediate layers of a deep CNN, with a simple curvy–boxy perceptual axis explaining a modest amount of the structure; and (*iv*) texform model features were able to account for a substantial amount of the neural similarity structure elicited by the original recognizable images.

Taken together, these findings establish that differences in mid-level features can drive extensive categorical neural responses along the ventral stream and underlie the topographic organization by animacy and object size. Broadly, these results inform the debate about the nature of object representation in occipitotemporal cortex: First, they challenge a simple conception of the visual hierarchy, as relatively primitive texforms drove category differences in what is typically considered high-level visual cortex. Second, they highlight that curvature covaries with broad category distinctions and provide an intuitive description of the kind of mid-level featural information represented in this cortex. Below, we discuss the implications of these findings for models of the ventral stream, whether purely low-level features could account for these findings, the role of curvature in ventral stream organization, and why we observe a gap in neural responses to texforms vs. original images.

Implications for Models of the Ventral Stream. There are two main observations to note about the texform topographies, each with separate implications for the nature of representation along the ventral stream. The first observation is that the neural differences between different kinds of texforms are detectable at all. Consider Fig. 1: These stimuli all look like textured blobs. Participants have no idea what they are seeing or even that there are different kinds of things here. One real possibility was that the differences between the texforms would be far too subtle to drive any measurable differences in brain responses, especially measured with fMRI. However, the data show that the visual system not only tracks these incoming texforms but also triggers specific neural responses that cleanly align with the animacy and the object size organizations. These data provide strong evidence that these regions do not require clearly defined features, such as eyes and tails or handles or even outer contours, to trigger responses that distinguish animals and objects of different sizes. Instead, these data provide evidence that a more statistical and primitive level of features supports broad category distinctions along the ventral stream.

The second observation is that these texform topographies actually extend much farther anterior than one might expect from a classic view of the ventral stream as a hierarchy. Within this classic view, neural regions require increasingly complex visual features to trigger a response (43). A widely held assumption is that the more complex identity-level representations in anterior regions achieve this more abstract and invariant level of representation at the expense of sensitivity to lower-level visual information such as visual field position, simple orientation, and spatial frequency (1, 43, 59, 60). Within this strict conceptualization of the hierarchy, texforms should solely drive differences in more posterior areas, e.g., those implicated in processing texture and curvature (61, 62), and not differences in more anterior regions, as they clearly lack the features that enable identity-level recognition. However, we found that texforms drive responses along the entire ventral stream. These empirical findings support a growing view in which higher-level visual cortex is sensitive to features that span multiple levels of representation (13): Anterior regions seem to retain some sensitivity to low- and mid-level features while also becoming increasingly tolerant of complex stimulus transformations.

Low-Level vs. Mid-level Features. We have argued for a mid-level of representation underlying occipitotemporal responses. However, could even lower-level features explain these results? Model comparison and neuroimaging data provide convergent evidence that simple low-level features are not sufficient to account for the animacy and object size activations along occipitotemporal cortex. First, we directly considered several low-level feature models, quantifying how well tuning along these features could predict the neural response structure in occipitotemporal cortex. These models performed poorly, especially relative to the more complex mid-level models (i.e., intermediate layer responses from a CNN; see also refs. 53 and 54). In fact, even the feature space we used to generate the texforms was unable to linearly predict the neural responses in occipitotemporal cortex, implying that the relevant visual features preserved in texforms are related to nonlinear combinations of the simpler texture-synthesis features. Second, we measured neural responses to texforms presented in upper and lower visual fields, finding that texforms still evoked an animacy and size organization that was tolerant of visual field position. This result argues against an account in which local, retinotopic, low-level feature tuning explains occipitotemporal responses.

Beyond these methods, another way to examine the contribution of low-level features in occipitotemporal cortex responses would be to create stimuli that only preserve relatively low-level image statistics. A recent study did something similar, using globally scrambled images, and found some correspondence between original images and their globally scrambled counterparts (37). However, it is likely that their specific analysis procedures led to somewhat biased results (38). Further, consistent with

the present results, they also found that a local-scrambling condition, which preserved coarse form and texture information, elicited activations that were much more similar to those of the original recognizable images than did the globally scrambled images. Taken together, these results suggest that while occipitotemporal responses may exhibit some tuning to very low-level features, a bulk of the response likely reflects tuning at a mid-level of representation, where the relative positions of local features matter.

The Role of Curvature in Ventral Stream Organization. We found that the similarity in the structure of neural responses across occipitotemporal cortex was well predicted not only by intermediate and later layers of a deep CNN but also by a single intuitive dimension of perceived curvature; this was true for both original images and unrecognizable texforms. This finding joins other research documenting the importance of curvature in ventral stream responses. For example, an elegant series of studies demonstrated the explanatory power of curvature in explaining single-unit responses in V4 (62–64). Other work has shown systematic preferences for curvilinear versus rectilinear stimuli in different category-selective regions in the inferior temporal cortex (34, 36, 50, 65, 66, but see ref. 67). Most recently, curvature has been proposed as a proto-organizing dimension of the ventral visual stream (50, 58), and specific curvature-preferring patches have been discovered in macaques (68). One challenge is that these studies have operationalized curvature in different ways (e.g., wavy-to-straight, round-to-rectilinear, curvy-to-boxy). Going forward, it will be important to develop a quantitative model that operationalizes curvature in a way that can unify these findings.

Why might curvature be such an important mid-level property? We have previously speculated there is an ecological (nonarbitrary) relationship between curvature and category: Big objects tend to be boxier because they must withstand gravity, while small objects tend to be curvier as they are made to be hand-held, and animals have few if any hard corners and are the curviest images (46, 47, 69, 70). In recent work, we have found direct evidence for this link (47, 48): The curviest texforms tend to be perceived as animate, and the boxiest texforms tend to be perceived as big, inanimate objects. Thus, the perceptual axis from boxy to curvy seems to align meaningfully with the broad category divisions between animals and objects of different sizes.

Based on these sets of results, we suggest that ventral stream responses are tuned according to mid-level feature maps that meaningfully covary with high-level, categorical dimensions. That is, the level of representation in the neural populations is visual/statistical in nature, but the organization of this feature tuning is still reasonably described by high-level animacy and object size distinctions. This work helps to refine our previous work showing that the high-level properties of object size and animacy distinctions yield a tripartite organization of the ventral stream (see direct comparison in *SI Appendix*, Fig. S11) (10, 11). Specifically, that the cortex is organized by these high-level factors does not mean that the nature of the tuning is also high-level—we think it is unlikely this cortex is directly computing an abstract sense of size or animacy per se. Rather, the present data support the idea that occipitotemporal cortex is largely computing visual shape structure, where animacy and object size are related to major axes through this shape space.

Of course, one of the big unanswered questions about the relationship between mid-level features and high-level organization is the direction of causality. Are broad category distinctions such as animacy and size evident because there are initial curvature biases in the visual system? For example, on an input-driven account, the statistics of visual experience with animals and objects of different sizes might be sufficient to account for this large-scale organization: Early retinotopy might naturally give rise to a large-scale curvature proto-organization in occipitotemporal cortex (11, 31, 50, 58) which in turn gives rise to a large-scale organization by the covarying high-level distinctions of animacy and object size. Alternatively, these mid-level curvature features might be learned specifically due to higher-level pressures to distinguish animals, big objects, and small objects (71, 72). For example, distinct whole-brain networks that support behaviors such as navigation, social interaction, and tool manipulation might specifically enforce animacy and object size shape-tuning in different regions. Note that the present data cannot speak to the directionality of these low-, mid-, and high-level factors but speak only to the existence of the link among them.

Differences Between Texform and Original Responses. While we have emphasized the extensiveness of the texform topographies, they are certainly distinguishable from the neural responses evoked by original, recognizable images. First, original images generated stronger categorical responses than texforms across the entire ventral stream in both the univariate effects and in their multivoxel patterns. Second, CNN features extracted in response to original images were necessary to best predict the neural structure generated by recognizable images; texform CNN features did well but did not reach the same level as the original image CNN features. What accounts for this gap between texform and original images?

It is tempting to consider attentional mechanisms as an explanatory factor; e.g., recognizable images could be more salient attentional stimuli than texforms, thereby driving stronger animacy/size preferences. However, it is important to note that CNN models were quite successful at predicting the structure of the occipitotemporal responses to both texforms and original images and also showed a gap between texforms and original images without relying on attentional mechanisms. Thus, texforms might instead drive weaker topographies because they are missing some critical visual features. What might these visual features be?

A first possibility is that original, recognizable images contain category-specific visual features that are captured by the CNN. For example, these category-specific features could include different sets of characteristic shape parts that differ between "animates and inanimates" [e.g., animals tend to have tails, eyes, and ears (70); small objects often have handles and buttons; big objects may have more extended flat surfaces]. A second possibility is that recognizable images contain additional generic visual features that are useful for describing any given object. For example, recognizable images contain strong bounded contours and other visual features that specify their 3D part-structure, whereas texforms do not. Thus, an alternative possibility is that these kinds of generic visual features, not tied to the category membership of the objects, account for this differential activity.

At stake in this distinction is whether the nature of the visual representation in occipitotemporal cortex should be considered more low level or high level (16, 55). Interestingly, CNNs might be able to provide some insight into these questions. For example, if a CNN were trained to perform a simpler task (e.g., a same vs. different image task), then the units would become tuned without any category-specific feedback but presumably would contain some set of generic visual descriptors. However, perhaps some degree of categorization training (e.g., animate/inanimate, or face/nonface) may be needed to render CNN units complex enough to predict categorical neural responses.

Conclusion

The present work investigated the link between mid-level features and the known animacy and size organizations in occipitotemporal cortex. We found that mid-level feature differences are sufficient to elicit these large-scale organizations along the entire ventral stream. Predictive modeling provided converging support for this result, as both intermediate layers of CNNs and intuitive ratings of curvature predicted the similarity in neural patterns elicited by texforms and recognizable images. This work provides evidence to situate the level of representation in the ventral stream, demonstrating that much of object-selective cortical organization can be explained by relatively primitive mid-level features without requiring explicit recognition of the objects themselves. Broadly,

these data are consistent with the view that the entire ventral stream is predominantly tuned to mid-level visual features, which are spatially organized across the cortex in a way that covaries with high-level, categorical dimensions.

Materials and Methods

Participants. Sixteen healthy observers (age range 18–35 y; seven females) with normal or corrected-to-normal vision participated in a 2-h fMRI session for experiment 1 (n = 8) and experiment 2 (n = 8). All participants (n = 110 across norming and fMRI experiments) provided informed consent. Procedures were approved by the Institutional Review Board at Harvard University.

Stimulus Set. The stimulus set consisted of 240 total images with 120 original images of 30 big animals, 30 big objects, 30 small animals, and 30 small objects and their texform counterparts.

Texforms were created using the following procedure. First, images were normalized for luminance and contrast across the whole set using the SHINE toolbox (73). Next, each image was placed in a larger gray display at a peripheral location so it fell within the larger spatial pooling windows generated by the texture synthesis algorithm (see *SI Appendix*, Fig. S1*A* for an illustration, as used in refs. 46–48). The synthesis algorithm proceeds by taking thousands of first- and second-order image-statistics measurements from the display, e.g., Gabor responses of different orientations, spatial frequencies, and spatial scales. Critically, however, these image statistics are computed within the local pooling windows (45), differentiating this method from previous texture-synthesis algorithms. Next, the algorithm starts with a white noise display and coerces the display to match the measured image statistics, using a variant of gradient descent that was terminated after 50 iterations. Then, online norming studies were conducted on a superset of 240 texforms to choose a set of unrecognizable texforms (*SI Appendix*, Fig. S2 *A* and *B*).

Texform Classifiability. The classifiability of each texform by its animacy/size was calculated using online rating experiments (*SI Appendix*, Fig. S3*A*). Specifically, one group of participants (n = 16) was shown a texform and was asked: "Here is a scrambled picture of something. Was the original thing an animal?" Participants responded "yes" or "no." Similarly, three other groups of participants (n = 16 in each group) judged whether the texform was a manmade object, was big enough to support a human, or was small enough to hold with one or two hands. Animacy and size classifiability scores were calculated for each image as the percent of correct classifications minus the percent of incorrect classifications. For example, if the texform was generated from an animal original image, this score was calculated as the percent of responses "Yes, it's an animal" minus the percent of responses "Yes, it's a manmade object." The same procedure was followed for size classifiability, e.g., the percent of answers "Yes, it's big" minus the percent of answers "Yes, it's small" if the original item had a big real-world size and the percent of answers "Yes, it's small" minus the percent of answers "Yes, it's big" if the item had a small real-world size. This serves as a proxy for a d-prime measure and allows response bias to be factored out from the classification scores. With these measures, the higher the score, the more the image was correctly classified as an animal or an object and as big or small; negative scores indicate systematic misclassifications. These animacy and size classification scores were summed to obtain a composite classification score which was used to assign the stimuli into six groups of five images per condition (big animals, big objects, small animals, small objects), from lowest to highest total classifiability (*SI Appendix*, Fig. S3*B*).

fMRI Experiment Design. Observers viewed images of big animals, small animals, big objects, and small objects in a standard blocked design while undergoing functional neuroimaging (*SI Appendix*). In the first four runs of the experiment, observers saw texforms; in the second four runs observers saw original images. Observers were not told anything regarding what the texforms were. Unknown to participants, the texform and original runs were yoked, such the original images were shown in exactly the same sequence and timing as the texforms. The observer's task was to pay attention to each item and to press a button when an exact image repeated back-to-back, which occurred once per block.

Preference Map Analyses. The spatial distribution and strength of response preferences in visually active voxels along the ventral stream were visualized using a preference-map analysis (10, 11). Active occipitotemporal cortex in each participant was defined to include all voxels with all conditions > rest with $t > 2$ in either texform or original runs, excluding voxels within the functionally defined early visual areas V1–V3 (*SI Appendix*, Fig. S6). For the animacy organization, for each voxel, the average beta for animals (across big and small sizes) was subtracted from the average beta for objects (across big and small sizes), and this beta-difference map was displayed on the cortical surface. For the size organization, for each voxel, the beta for big objects was subtracted from the beta for small objects and was displayed on the cortical surface. To compare animacy and real-world size-preference maps elicited by texform and original images, we used a map-correlation procedure following ref. 50. The map correlation was computed as the correlation over voxels between the beta-difference scores for the texform organization and original organization and was computed separately for each subject for both animacy and object size dimensions. See *SI Appendix* for details on the shuffled baseline and noise ceiling calculations.

Posterior-to-Anterior Analyses. In each participant, anatomical sections were defined along a posterior-to-anterior gradient within occipitotemporal cortex by dividing it into five quantiles using the TAL-Y coordinates of visually active voxels [taken from each participant's generalized linear model (GLM) data]. A measure of the strength of the animacy (size) preferences for either objects or texforms was computed as the absolute value of animals vs. object betas (big vs. small object betas) for each voxel, averaged across voxels. These estimates were computed separately for original images and texforms in each section and in each participant. See *SI Appendix*.

Conjunction Analysis. Conjunction voxels were defined as those that elicited the same category preference (e.g., animals) regardless of the location of the image in the visual field (i.e., the upper visual field or the lower visual field) in response to recognizable images (i.e., the original images). Conjunction voxels were defined separately within each subject (*SI Appendix*, Fig. S13). To calculate the portion of retained voxels, we divided the number of voxels in this conjunction mask by the total number of visually active voxels in occipitotemporal cortex in each subject. Map correlations were then performed in each subject within these conjunction voxels.

Representational Similarity Analysis. Multivoxel patterns were extracted for each of the four main conditions (animals/objects × big/small) at each level of classifiability (levels 1–6), yielding 24 conditions. Given that each voxel is treated as a separate dimension in this analysis, we considered only voxels where recognizable images yielded a split-half reliability value above zero (*SI Appendix*, Fig. S16). Next, the correlation distance between neural patterns within these voxels was computed separately for texforms and original images for each participant and was averaged for the group visualization of the RDM.

Predictive Modeling Approach. To compare how well different models (i.e., low-level feature models, CNN features, and behavioral ratings) could predict the neural RDMs, we used weighted representational similarity analysis (RSA), a predictive modeling procedure (51). First, for each model, features were extracted from each image in the set and were averaged by classifiability group into a 24-condition × numFeature matrix (see *SI Appendix* for more details on feature extraction). Next, each feature was converted from a 24 × 1 vector into a vectorized RDM (276 × 1), in which the 276 values correspond to the squared Euclidean distance between all possible pairs of 24 conditions. Here, vectorized RDMs reflect only the values in the upper triangle of the matrix, excluding the diagonal. Using nonnegative least squares regression (lsqnonneg in Matlab 2015a), we modeled the brain-vectorized RDM as a weighted combinations of these feature-vectorized RDMs, with a leave-one-condition-out cross-validation procedure. In each iteration, one of the 24 conditions was dropped from both the brain and feature data, removing 23 cells from the 276 vector (e.g., dropping condition 1 removes the distances between conditions 1 and 2, between conditions 1 and 3, between conditions 1 and 4, and so forth). The fitted model weights were then used to predict the distance to these held-out points. A predicted RDM was compiled over all cross-validation iterations in which the two predictions for the same dissimilarities pairs were averaged (e.g., the distances from conditions 1 and 2 and from conditions 2 and 1) to make it symmetric. The goodness-of-prediction was assessed by correlating the predicted vectorized RDM with each subject's neural vectorized RDM. This procedure was employed for all different feature models. Finally, the noise ceiling of the neural data was computed using the RSA toolbox (52) reflecting the degree to which an individual subject's RDM could predict the group's RDM.

1. DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. *Trends Cogn Sci* 11: 333–341.
2. Mishkin M, Ungerleider LG, Macko KA (1983) Object vision and spatial vision: Two cortical pathways. *Trends Neurosci* 6:414–417.
3. Cohen L, et al. (2000) 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.
4. Downing PE, Jiang Y, Shuman M, Kanwisher N (2001) A cortical area selective for visual processing of the human body. *Science* 293:2470–2473.
5. Epstein R, Kanwisher N (1998) A cortical representation of the local visual environment. *Nature* 392:598–601.
6. Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J Neurosci* 17:4302–4311.
7. Haxby JV, et al. (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293:2425–2425.
8. Julian JB, Ryan J, Epstein RA (2017) Coding of object size and object category in human visual cortex. *Cereb Cortex* 27:3095–3109.
9. Chao LL, Haxby JV, Martin A (1999) Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nat Neurosci* 2:913–919.
10. Konkle T, Caramazza A (2013) Tripartite organization of the ventral stream by animacy and object size. *J Neurosci* 33:10235–10242.
11. Konkle T, Oliva A (2012) A real-world size organization of object responses in occipitotemporal cortex. *Neuron* 74:1114–1124.
12. Martin A (2007) The representation of object concepts in the brain. *Annu Rev Psychol* 58:25–45.
13. Grill-Spector K, Weiner KS (2014) The functional architecture of the ventral temporal cortex and its role in categorization. *Nat Rev Neurosci* 15:536–548.
14. Kravitz DJ, Vinson LD, Baker CI (2008) How position dependent is visual object recognition? *Trends Cogn Sci* 12:114–122.
15. Lambon Ralph MA, Jefferies E, Patterson K, Rogers TT (2017) The neural and computational bases of semantic cognition. *Nat Rev Neurosci* 18:42–55.
16. Peelen MV, Downing PE (2017) Category selectivity in human visual cortex: Beyond visual object recognition. *Neuropsychologia* 105:177–183.
17. Castelli F, Happé F, Frith U, Frith C (2000) Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage* 12:314–325.
18. Wheatley T, Milleville SC, Martin A (2007) Understanding animate agents: Distinct roles for the social network and mirror system. *Psychol Sci* 18:469–474.
19. Weisberg J, van Turennout M, Martin A (2007) A neural system for learning about object function. *Cereb Cortex* 17:513–521.
20. Bracci S, Op de Beeck H (2016) Dissociations and associations between shape and category representations in the two visual pathways. *J Neurosci* 36:432–444.
21. Kaiser D, Azzalini DC, Peelen MV (2016) Shape-independent object category responses revealed by MEG and fMRI decoding. *J Neurophysiol* 115:2246–2250.
22. Proklova D, Kaiser D, Peelen MV (2016) Disentangling representations of object shape and object category in human visual cortex: The animate-inanimate distinction. *J Cogn Neurosci* 28:680–692.
23. Ratan Murty NA, Pramod RT (2016) To what extent does global shape influence category representation in the brain? *J Neurosci* 36:4149–4151.
24. He C, et al. (2013) Selectivity for large nonmanipulable objects in scene-selective visual cortex does not require visual experience. *Neuroimage* 79:1–9.
25. Peelen MV, He C, Han Z, Caramazza A, Bi Y (2014) Nonvisual and visual object shape representations in occipitotemporal cortex: Evidence from congenitally blind and sighted adults. *J Neurosci* 34:163–170.
26. Striem-Amit E, Amedi A (2014) Visual cortex extrastriate body-selective area activation in congenitally blind people "seeing" by using sounds. *Curr Biol* 24:687–692.
27. van den Hurk J, Van Baelen M, Op de Beeck HP (2017) Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proc Natl Acad Sci USA* 114:E4501–E4510.
28. Bi Y, Wang X, Caramazza A (2016) Object domain and modality in the ventral visual pathway. *Trends Cogn Sci* 20:282–290.
29. Wandell BA, Dumoulin SO, Brewer AA (2007) Visual field maps in human cortex. *Neuron* 56:366–383.
30. Golomb JD, Kanwisher N (2012) Higher level visual cortex represents retinotopic, not spatiotopic, object location. *Cereb Cortex* 22:2794–2810.
31. Hasson U, Levy I, Behrmann M, Hendler T, Malach R (2002) Eccentricity bias as an organizing principle for human high-order object areas. *Neuron* 34:479–490.
32. Larson AM, Loschky LC (2009) The contributions of central versus peripheral vision to scene gist recognition. *J Vis* 9:6.1–16.
33. Levy I, Hasson U, Avidan G, Hendler T, Malach R (2001) Center-periphery organization of human object areas. *Nat Neurosci* 4:533–539.
34. Rajimehr R, Bilenko NY, Vanduffel W, Tootell RBH (2014) Retinotopy versus face selectivity in macaque visual cortex. *J Cogn Neurosci* 26:2691–2700.
35. Baldassi C, et al. (2013) Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons. *PLOS Comput Biol* 9:e1003167.
36. Nasr S, Echavarria CE, Tootell RBH (2014) Thinking outside the box: Rectilinear shapes selectively activate scene-selective cortex. *J Neurosci* 34:6721–6735.
37. Coggan DD, Liu W, Baker DH, Andrews TJ (2016) Category-selective patterns of neural response in the ventral visual pathway in the absence of categorical information. *Neuroimage* 135:107–114.
38. Ritchie JB, Bracci S, Op de Beeck H (2017) Avoiding illusory effects in representational similarity analysis: What (not) to do with the diagonal. *Neuroimage* 148:197–200.
39. Andrews TJ, Clarke A, Pell P, Hartley T (2010) Selectivity for low-level features of objects in the human ventral stream. *Neuroimage* 49:703–711.
40. Op de Beeck HP, Torfs K, Wagemans J (2008) Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *J Neurosci* 28:10111–10123.
41. Lerner Y, Harel M, Malach R (2004) Rapid completion effects in human high-order visual areas. *Neuroimage* 21:516–526.
42. Tanaka K (2003) Columns for complex visual object features in the inferotemporal cortex: Clustering of cells with similar but slightly different stimulus selectivities. *Cereb Cortex* 13:90–99.
43. Lehky SR, Tanaka K (2016) Neural representation for object recognition in inferotemporal cortex. *Curr Opin Neurobiol* 37:23–35.
44. Biederman I (1987) Recognition-by-components: A theory of human image understanding. *Psychol Rev* 94:115–147.
45. Freeman J, Simoncelli EP (2011) Metamers of the ventral stream. *Nat Neurosci* 14: 1195–1201.
46. Long B, Konkle T, Cohen MA, Alvarez GA (2016) Mid-level perceptual features distinguish objects of different real-world sizes. *J Exp Psychol Gen* 145:95–109.
47. Long B, Störmer VS, Alvarez GA (2017) Mid-level perceptual features contain early cues to animacy. *J Vis* 17:20.
48. Long B, Konkle T (2017) A familiar-size Stroop effect in the absence of basic-level recognition. *Cognition* 168:234–242.
49. Krizhevsky A, Sutskever I, Hinton G (2012) Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing System*, pp 1097–1105.
50. Srihasam K, Vincent JL, Livingstone MS (2014) Novel domain formation reveals proto-architecture in inferotemporal cortex. *Nat Neurosci* 17:1776–1783.
51. Jozwik KM, Kriegeskorte N, Mur M (2016) Visual features as stepping stones toward semantics: Explaining object similarity in IT and perception with non-negative least squares. *Neuropsychologia* 83:201–226.
52. Nili H, et al. (2014) A toolbox for representational similarity analysis. *PLOS Comput Biol* 10:e1003553.
53. Khaligh-Razavi SM, Kriegeskorte N (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Comput Biol* 10:e1003915.
54. Yamins DLK, et al. (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci USA* 111:8619–8624.
55. Andrews TJ, Watson DM, Rice GE, Hartley T (2015) Low-level properties of natural images predict topographic patterns of neural response in the ventral visual pathway. *J Vis* 15:3.
56. Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int J Comput Vis* 42:145–175.
57. Güçlü U, van Gerven MAJ (2015) Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J Neurosci* 35: 10005–10014.
58. Ponce CR, Hartmann TS, Livingstone MS (2017) End-stopping predicts curvature tuning along the ventral stream. *J Neurosci* 37:648–659.
59. Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex* 1:1–47.
60. Rust NC, Dicarlo JJ (2010) Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *J Neurosci* 30:12978–12995.
61. Kourtzi Z, Connor CE (2011) Neural representations for object perception: Structure, category, and adaptive coding. *Annu Rev Neurosci* 34:45–67.
62. Pasupathy A, Connor CE (2001) Shape representation in area V4: Position-specific tuning for boundary conformation. *J Neurophysiol* 86:2505–2519.
63. Carlson ET, Rasquinha RJ, Zhang K, Connor CE (2011) A sparse object coding scheme in area V4. *Curr Biol* 21:288–293.
64. Yau JM, Pasupathy A, Brincat SL, Connor CE (2013) Curvature processing dynamics in macaque area V4. *Cereb Cortex* 23:198–209.
65. Caldara R, et al. (2006) The fusiform face area is tuned for curvilinear patterns with more high-contrasted elements in the upper part. *Neuroimage* 31:313–319.
66. Rajimehr R, Devaney KJ, Bilenko NY, Young JC, Tootell RBH (2011) The "parahippocampal place area" responds preferentially to high spatial frequencies in humans and monkeys. *PLoS Biol* 9:e1000608.
67. Bryan PB, Julian JB, Epstein RA (2016) Rectilinear edge selectivity is insufficient to explain the category selectivity of the parahippocampal place area. *Front Hum Neurosci* 10:137.
68. Yue X, Pourladian IS, Tootell RBH, Ungerleider LG (2014) Curvature-processing network in macaque visual cortex. *Proc Natl Acad Sci USA* 111:E3467–E3475.
69. Konkle T, Oliva A (2011) Canonical visual size for real-world objects. *J Exp Psychol Hum Percept Perform* 37:23–37.
70. Levin DT, Takarae Y, Miner AG, Keil F (2001) Efficient visual search by category: Specifying the features that mark the difference between artifacts and animals in preattentive vision. *Percept Psychophys* 63:676–697.
71. Konkle T, Caramazza A (2016) The large-scale organization of object-responsive cortex is reflected in resting-state network architecture. *Cereb Cortex* 31:1–13.
72. Mahon BZ, Caramazza A (2011) What drives the organization of object knowledge in the brain? *Trends Cogn Sci* 15:97–103.
73. Willenbockel V, et al. (2010) Controlling low-level image properties: The SHINE toolbox. *Behav Res Methods* 42:671–684.