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

Bria Long[1,2*], Chen-Ping Yu[1,3], & Talia Konkle[1]

[1]Department of Psychology, Harvard University
[2]Department of Psychology, Stanford University
[3]Phiar Technologies, Inc.
* corresponding author

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 novel 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.

Significance Statement

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 years. In this paper, a new 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.

Introduction

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 the occipitotemporal cortex (3–6). Other basic-level category distinctions (e.g., shoes vs. keys) lack clear category-specific regions, yet they can also be decoded from multi-voxel patterns in this same cortex (7, 8). Even broader categorical distinctions reflecting the animacy and real-world size of objects are evident in 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 whatsoever (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). But, for the neural representations in intermediate occipitotemporal cortex, there is active debate about just how "high" or "low" the nature of the representation is.

On 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 (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 versus a rope; (20–22) but see (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, like 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 (38, but see 39). 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).

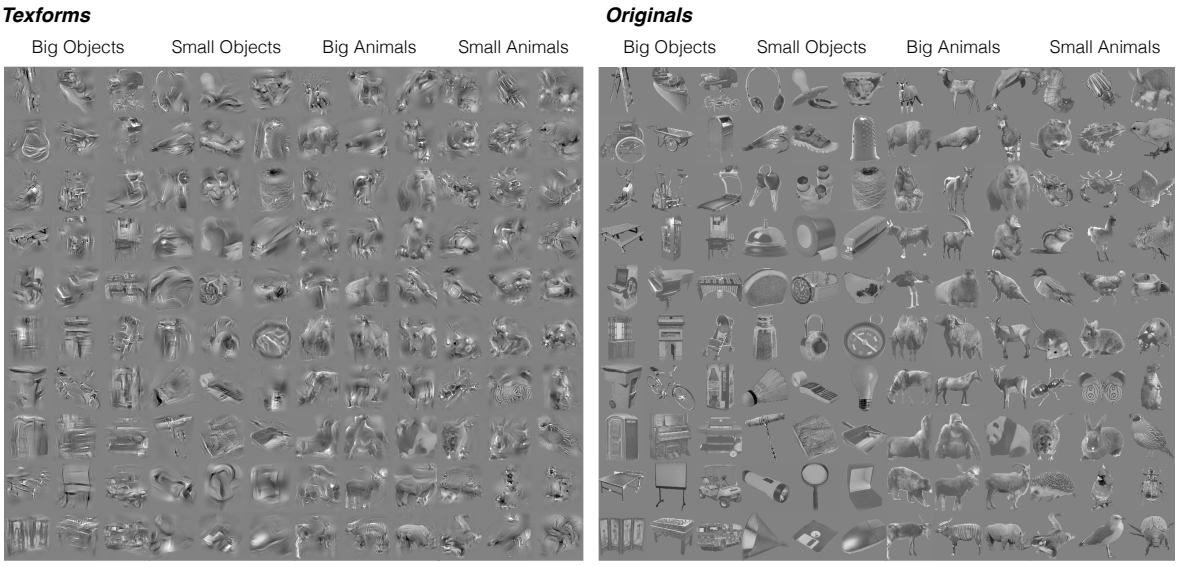

Figure 1. Stimuli: 120 texforms (right panel) were generated using a texture synthesis model (Freeman & Simoncelli, 2011) from recognizable pictures of 30 big objects, 30 small objects, 30 big animals, and 30 small animals (left panel). Stimuli are shown at slightly higher contrast for visualization purposes. Stimuli were selected such that all texforms were unrecognizable at the basic level using online recognition experiments.

These two current viewpoints represent two prominent models of how to characterize the representations in occipitotemporal cortex. On an intermediate account, of course, neural responses in occipitotemporal cortex reflect tuning to visual features of intermediate complexity (e.g., 13, 42–44). 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 feature space derived from generic building blocks (i.e. (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 is both methodologically and theoretically challenging.

Here, we approached this challenge by leveraging a new 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 look to most people like "texturey-blobs" (see **Figure 1**; **Figure S1**; (45–48). These stimuli have two properties that make them particularly well-suited to probe 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; **Figure 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 (e.g. **Figure S3; Figure 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 (49) were able to explain a substantial portion of the variance in neural pattern to both 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 occipito-temporal 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 co-vary with high-level distinctions between object categories, and 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. **Figure 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 (**Figure S3**). These scores were used to systematically vary how classifiable each block of texforms was (see 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 multi-voxel 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 **Figure S5).**

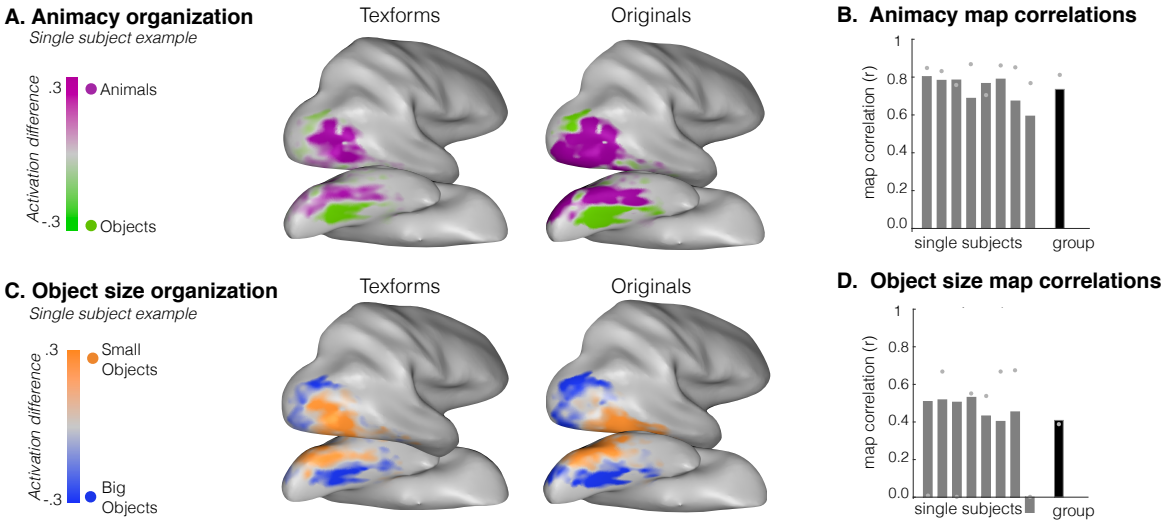

Figure 2. Preference map analyses. Response preferences within active occipitotemporal voxels are plotted for (A) animals versus objects and (C) big vs. small objects for an example participant, considering texform images (left), and original images (right). The color bar reaches full saturation at activation differences between .3 and -.3 (reflecting the beta difference calculated from this individual's GLM). The correlation between the original and texform response maps in active occipitotemporal voxels is plotted for animacy (B) and object size (D) distinction. Original-texform map correlations are shown for all individual participants, and the group, averaged across all subjects. Grey dots indicate the estimated noise ceiling for each participant and at the group level.

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, $t > 2$; **Figure S6**). Systematic response differences to animal versus object texforms were observed across the entire occipitotemporal cortex, with a large-scale organization in their spatial distribution. The 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 **Figure 2a** and reveal a striking degree of similarity (see group topographies in **Figure S7;** all single subject topographies in **Figure 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 (50) (see Methods). The map correlation coefficients for each participant and the average correlation coefficient for the group are plotted in **Figure 2b**. Overall, voxels in the occipitotemporal cortex had similar animacy preferences for recognizable and texform images in all subjects, resulting in a robust correlation at the group level (Average $r = .74$, $SD = .07$, permutation test significant in each subject, all $p < .001$; average noise ceiling across subjects, r

= .81, *SD* = .06, see Methods).

Next, we used a similar analysis to examine whether texforms also elicited a real-world size organization. Given that the size organization is only found 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 **(Figure 2c**; see group topographies in **Figure S7** and all single-subject topographies in **Figure 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 (**Figure 2d**, Average $r = .41$, $SD = .20$, permutation test significant in 7/8 subjects at $p < .001$, see Methods). 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 (see **Figure 2d**, average noise ceiling across subjects, $r = .39$, $SD = .32$).

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 originals? To examine this possibility, we split the data in half by texform classifiability. For the animacy distinction, map correlations between originals and texforms were higher for better classified texforms ($M = .73$) than for worse classified texforms ($M = .45$, $t(7) = 7.00$, $p < .001$). However, the size organization was not as strongly influenced by classifiability (average map correlation for better classified texforms vs. originals, $M = .35$; worse classified texforms vs. originals $M = .29$; $t(7) = 1.25$, $p = 0.25$).

There are at least two possible accounts of 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 while 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 (see Methods; **Figure 3**).

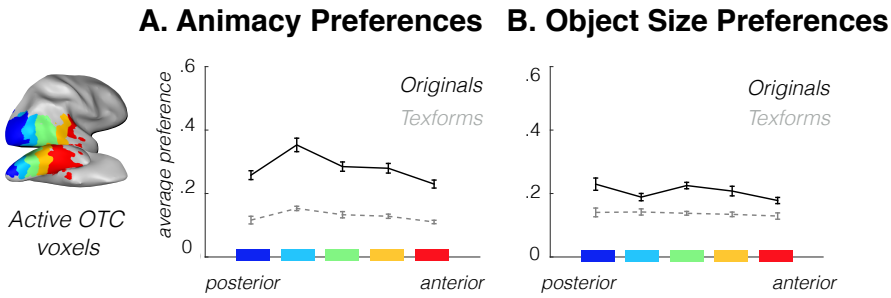

Figure 3. Anatomical sections (shown here at the group level) from posterior to anterior in blue to red. (A) The strength of the average animal/object response preference is shown for each anatomical section, averaged across voxels and participants, plotted for both originals (black solid line) and texforms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

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. The overall strength of the animal/object preferences in each sector is shown in **Figure 3a** (see also **Figure S9)**, where the solid lines show the average preference strength for originals, and the dashed lines 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 (**Figure 3a**; average activation difference for each section in animacy preferences for original - texforms: M_{s1}= .14, M_{s2}= .20, M_{s3}= .15, M_{s4}=.15, M_{s5} = .12; average rank correlation across subjects between these activation differences and anatomical sections, r = -.34, t-test against zero, t(7) = -1.92, p = 0.10). Thus, this analysis reveals that original images generate stronger category preferences than do texforms across *all* anatomical sections, and 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 (**Figure 3b**), average activation difference for each section in object size preferences for originals - texforms: M_{s1}= .09, M_{s2}= .05, M_{s3}= .09, M_{s4}= .07, M_{s5} = .05; average rank correlation between these activation differences and anatomical sections, r = -.25, t-test against zero, $t(7)$ = -1.17, p = 0.28). 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 to originals versus texforms remained relatively constant across the full posterior-to-anterior gradient.

One possible factor that might influence the interpretation of this result is about 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 reflects better signal-to-noise. On such an account, 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 originals 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, though if anything, recognizable images generated slightly more overall activity than texforms in the more *anterior* sections (average activation difference between originals and texforms in each section; M_{S1} = -0.01, M_{S2} = .07, M_{S3} = .12, M_{S4} = .11, M_{S5} = .14, average rank correlation between activation differences and anatomical sections, r = .59, t-test against zero, $t(7) = 3.84$, $p = 0.006$, see **Figure S10**). Thus, it was not the case that original images elicited stronger overall responses everywhere, and 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 (see also **Figure S11**). We did not find strong evidence for a hierarchy of representations that differentiated between texforms and recognizable images. Instead, these results point towards mid-level features as a major explanatory factor for the spatial topography of object responses along the entire occipitotemporal cortex.

Tolerance to 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 new group of observers were shown the same stimuli (both texforms and recognizable images) but each image was presented separately above and below fixation (**Figure S12**). If animacy and size preferences are maintained over changes in visual field position, this provides evidence against a simple retinotopic account.

In our first analysis, we examined how much of the occipitotemporal cortex showed location-tolerant animacy and size preferences, separately for originals 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 (see **Figure 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.

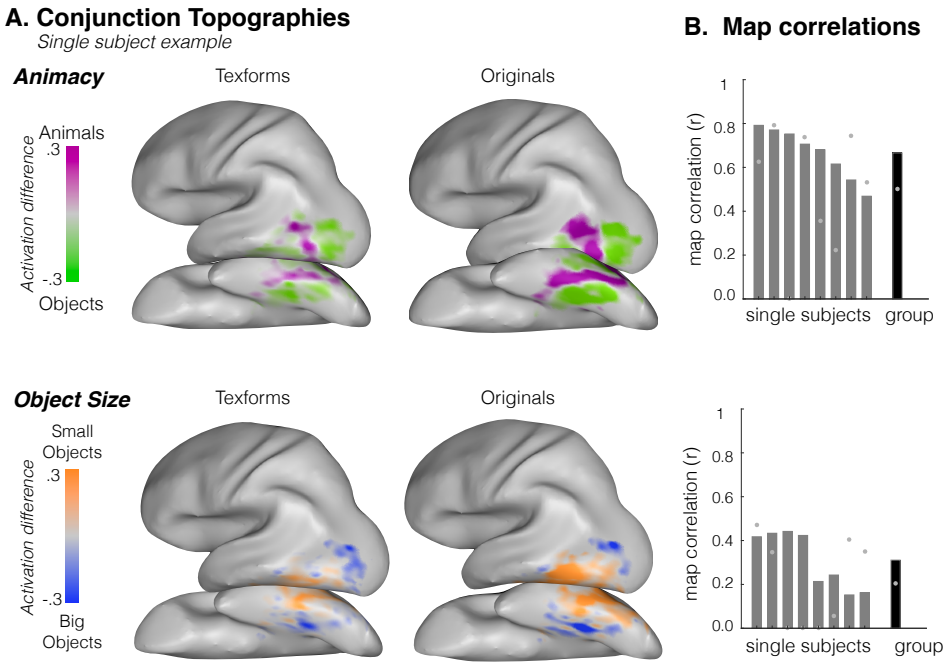

Figure 4. (A). Group conjunction topographies. Average category responses when texforms (left) and originals (right) are presented in the upper and lower visual field. Topographies are restricted to voxels that show the same category preference regardless of the stimuli's location in the visual field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

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 originals 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/originals are shown within these location-tolerant, conjunction voxels in **Figure 4a**, and qualitatively show similar spatial profiles (see group topographies in **Figure S14** and all single subject topographies in **Figure 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 = .67, *SD* = .12, Size: average r = . 31, *SD* = .11, permutation tests against shuffled voxel baseline significant in all subjects at $p < .001$; see **Figure 4b**).

Furthermore, when we relaxed our voxel inclusion criterion – analyzing map correlations within *all* visually-active voxels in OTC, as in Experiment 1, we found the same pattern of results (Animacy: average r = .60, SD = .11, Size: average r = . 25, SD = .11), indicating that the stringent voxel inclusion criteria did not bias the results.

Compared to 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 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 to 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 multi-voxel 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 (**Figure 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 were able to guess that this texform was generated from an animal vs. an object, or from a big vs. small thing in the world (**Figure S3**). Examples of well-classified and poorly-classified texforms (and their accompanying original counterparts) are shown in **Figure 5a**.

Figure 5b shows the similarity in the multivoxel patterns elicited by texforms, and the corresponding originals. The texform RDM has some gradation in terms of the levels of texforms' 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 convolutional neural network (49), and behavioral ratings of perceived curvature, using a weighted representational modeling procedure as introduced by (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 (see Methods, (51)) and 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 (see Methods, (52)) and shown in **Figure 6a**.

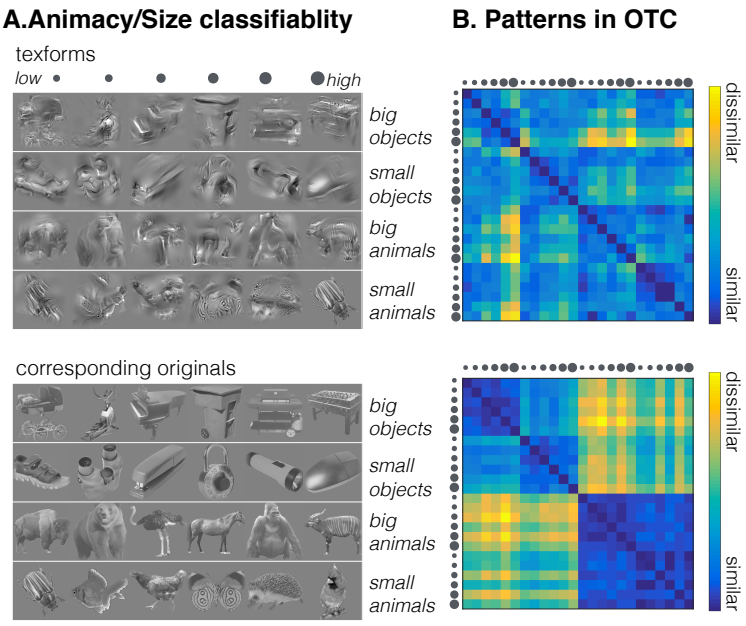

Figure 5. (A) An example texform is shown from the 6 classifiability groups, from lowest to highest, for the 4 main conditions, with the corresponding original images below. (B). Representational dissimilarity matrix obtained from neural patterns in active occipitotemporal cortex for texforms (top) and originals (bottom). Data are scaled such that, in both cases, the most dissimilar values are yellow and the least dissimilar values are blue.

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 insufficient basis for predicting the geometrical layout of categorical responses in IT 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 (τ_A= .16; noise ceiling τ_A= [.38-.48]). Consistent with this result, other models based on low-level image statistics also only captured a small amount of variance (Gabor model, τ_A= .12; Gist model, τ_A= .10; (56)). Thus, linear combinations of relatively simple visual features were not sufficient to predict the multi-voxel patterns of occipitotemporal cortex.

Convolutional Neural Network (CNN). 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, see SI) in response to texforms. Note that while this CNN was pre-trained to categorize one thousand object categories, it was not specifically trained (or tuned) on 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, τ_A =.31; Layer 5, τ_A =.32). Thus, the variation in neural patterns 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 to different groups of texforms.

Curvature Ratings. We next asked how well perceived curvature ratings could explain this neural structure, based on behavioral evidence that boxy/curvy ratings distinguish between animals, small objects, and big objects (46–48) (see **Figure S4**), 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 (τ_A =.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, τ_A =.38; noise ceiling τ_A = [.38—.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 **Figure 6a**. To visually inspect the structure captured by the different models, predicted neural RDMs are shown from several models. 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 to 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. And, critically, the CNN does not have top-down feedback, and thus no mechanism by which to amplify any animacy/size differences (see also **Figure S17, Figure S18** for modeling results in each anatomical section of OTC). Thus, the modeling results are consistent with the idea that some texforms are more classifiable than others because they preserved more of the relevant mid-level features.

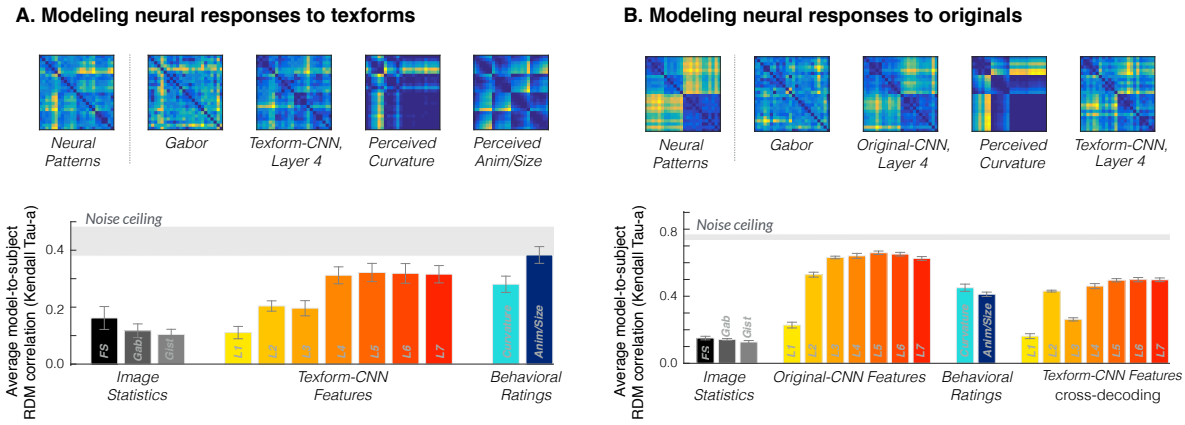

Figure 6. (A) Neural patterns in response to texforms (same as in Figure 5B) and predicted neural dissimilarities for selected models obtained through the same cross-validation procedure. The bar plot shows the predicted model correlation (Kendall Tau-a), where the error bars reflect the standard error of the model fit across individual subjects' neural patterns in OTC. 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 layer 1 through layer 7 (yellow to red), curvature behavioral ratings (light blue), and animacy/size behavioral ratings (dark blue). Data is plotted with respect to the noise ceiling of neural responses to texform images across participants, shown in light gray. (B) Neural patterns in response to original images (same as in Figure 5C) and predicted neural dissimilarities for four models obtained through the same leave-one-out cross-validation procedure. The average predicted model correlation (Kendall Tau-a) is plotted for different models, as in (A), with AlexNet features extracted from both original images and texforms. Data is plotted with respect to the noise ceiling of neural responses to original images across participants, shown in light gray.

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, with the results summarized in **Figure 6b** (see also **Figure S19).** Overall, we found a similar pattern of results as we did with texforms: basic image statistic models performed poorly (Freeman & Simoncelli features, τ_A =.15; Gabor features, τ_A= .14; Gist model, τ_A= .13; Neural noise ceiling τ_A= [.73—.77]), while the feature representations elicited in deep CNNs by recognizable images almost fully predicted these neural patterns by intermediate layers (Layer 4, τ_A= .64; Layer 5, τ_A= .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 (τ_A= .45). Finally, a categorical model of animacy and object size only performed moderately well (τ_A= .41), consistent with prior work highlighting that OTC 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, texform-CNN-features were able to predict much of the RDM structure elicited by recognizable images (Layer 4, T_A= .46, **Figure 6b**). 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 novel 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 more high-level object-selective areas, (ii) these mid-level topographies were not inherited from low-level retinotopic biases, as they generalized over visual field position, (iii) the similarity structure of the neural representations to 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 mid-level feature differences 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 co-varies 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. originals.

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 Figure 1—these stimuli all look like texturey-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 is not only tracking these incoming texforms, but it is also triggering 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, like eyes and tails or handles or even outer contours, in order 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 support broad category distinctions along the ventral stream.

The second observation is that these texform topographies actually extend much farther anterior than one might expect on a classic view of the ventral stream as a hierarchy. Within this classic view, neural regions require increasingly complex visual features in order to even 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 like 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 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 to complex stimulus transformations.

Low-level vs. mid-level features. We have argued for a mid-level of representation underlying occipitotemporal responses. But, could even lower-level features explain these results? Both 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 OTC. These models performed poorly, especially relative to the more complex "mid-level" models (i.e., intermediate layer responses from a CNN; see also (53, 54)). In fact, even the feature space we used to generate the texforms was unable to linearly predict the neural responses in occipito-temporal cortex, implying that the relevant visual features preserved in texforms are related to non-linear 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 to visual field position. This result argues against an account where 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 lead 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 much more similar activations to 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 structure of neural responses across occipitotemporal cortex was not only well predicted by intermediate and later layers of a deep convolutional neural network, 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 infereo-temporal cortex (36, 50, 65–67) but see (68). 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 (69). 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 (non-arbitrary) 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 (46, 47, 70, 71). 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 meaningfully align 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 co-vary 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 (10, 11); see direct comparison in **Figure S11**). 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 like 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 co-varying 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 (72, 73). For example, distinct whole brain networks that support behaviors like 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, only to the existence of the link between 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 multi-voxel 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-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 originals, and also showed a gap between texforms and originals, 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. These category-specific features could include, for example, different sets of characteristic shape parts that differ between animates and inanimates (e.g., animals tend to have tails, eyes, ears (71); 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, on an alternative account, it is these kinds of generic visual features, not tied to the category membership of the objects, which account for this differential activity.

At stake with 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, convolutional neural networks might be able to provide some insight into these questions. For example, if a CNN was trained to perform a simpler task (e.g., a same vs. different image task), then their units would become tuned without any category-specific feedback, but would presumably contain some set of generic visual descriptors. However, perhaps some degree of categorization training (e.g., animate/inanimate, or face/non-face) 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 neural pattern similarity elicited by texforms and recognizable images. This work provides new 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 co-varies with high-level, categorical dimensions.

Materials and Methods

Participants. Sixteen healthy observers with normal or corrected-to-normal vision participated in a 2-hour fMRI session (age range 18–35 years; seven females) for Experiment 1 ($N = 8$) and Experiment 2 ($N = 8$). All participants ($N = 110$ across norming and fMRI experiments) were consented using procedures 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, as well as 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 (74). Next, each image was placed in a larger gray "display" at a peripheral location, so as to fall in the larger spatial pooling windows generated by the texture synthesis algorithm (see **Figure S1A** for an illustration, as used in (46, 47)). The synthesis algorithm proceeds by taking thousands of first and second order image statistic 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 it to have match the measured image statistics, using a variant of gradient descent, terminated after 50 iterations. Then, online norming studies were conducted on a superset of 240 texforms to choose a set of unrecognizable texforms (see SI, **Figure S2A,B**).

Texform Classifiability. The classifiability of each texform by its animacy/size was calculated using online rating experiments (see **Figure S3A**). Specifically, one group of participants ($N = 16$) were shown a texform and asked: "Here is a scrambled picture of something. Was the original thing an animal?", and responded with "Yes" or "No". Similarly, three other groups of participants ($N = 16$ each) judged whether the texform was a "man-made object", "big enough to support a human", and "small enough to hold with one or two hands". Animacy and size classifiabilty scores were calculated for each image as %correct – %incorrect classifications. For example, if the texform was generated from an animal original image, this score was calculated as %(yes, was an animal) – %(yes, it's a man-made object). The same procedure was followed for size classifiability: %(yes, it's big) – %(yes, it's small) if the original item had a big real-world size; and %(yes, it's small) – %(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 for response bias to be factored out from the classification scores. With these measures, the higher the score, the more it was correctly classified as an animal/object and big/small; negative scores indicate systematic misclassifications. These animacy and size classification scores were summed to obtain a composite classification score, which were used to assign the stimuli into 6 groups of 5 images per condition (big animals, big objects, small animals, small objects), from lowest to highest total classifiability (see **Figure S3B)**.

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 (see SI). 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. Unbeknownst to participants, the texform and original runs were yoked, such the original images were shown in the exact same sequence and timing as the texforms. Observers' 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 was visualized using a preference-map analysis (10, 11). Active occipitotemporal cortex were defined in each participant to include all voxels with all conditions > rest with $t > 2$ in either texform or original runs, excluding voxels within functionally-defined early visual areas V1-V3 (see **Figure 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 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 (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 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 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-object betas (big-small object betas) for each voxel, averaged across voxels. These estimates were computed separately for originals and texforms, in each section, in each participant. See SI.

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., upper visual field, lower visual field) in response to recognizable images (i.e., originals). Conjunction voxels were defined separately within each subject (see **Figure 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 OTC in each subject. Map correlations were then performed in each subject within these conjunction voxels.

Representational Similarity Analysis. Multi-voxel patterns were extracted for each of the four main conditions (animals/objects x 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 only considered voxels where recognizable images yielded a split-half reliability value above zero (see **Figure S16**). Next, the correlation distance between neural patterns within these voxels was computed separately for texforms and original images, for each participant, and averaged for the group visualization of the representational dissimilarity matrix (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 RSA, a predictive modeling procedure (51). First, for each model, features were extracted from each image in the set, and averaged by classifiability group into a 24 condition x numFeature matrix (see SI for more detail on feature extraction). Next, each feature was converted from a 24x1 vector into a vectorized RDM (276 x 1), where 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 non-negative 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, 1 out 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 condition 1 to 2, condition 1 to 3, condition 1 to 4, etc.). 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, where the two predictions for the same dissimilarities pairs were averaged (e.g., the distance from condition 1 to 2 and the distance from condition 2 to 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.

References:

1. DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. *Trends Cogn Sci* 11(8):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(2):291–307.
4. Downing PE, Jiang Y, Shuman M, Kanwisher N (2001) A cortical area selective for visual processing of the human body. *Science (80-)* 293(5539):2470–2473.
5. Epstein R, Kanwisher N (1998) A cortical representation of the local visual environment. *Nature* 392(6676):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(11):4302–4311.
7. Haxby J V., et al. (2001) Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science (80-)* 293(5539):2425–2425.
8. Julian JB, Ryan J, Epstein RA (2016) Coding of Object Size and Object Category in Human Visual Cortex. *Cereb Cortex* bwh150:1–15.
9. Chao LL, Haxby J V, Martin A (1999) Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nat Neurosci* 2(10):913–919.
10. Konkle T, Caramazza A (2013) Tripartite organization of the ventral stream by animacy and object size. *J Neurosci* 33(25):10235–42.
11. Konkle T, Oliva A (2012) A Real-World Size Organization of Object Responses in Occipitotemporal Cortex. *Neuron* 74(6):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(8):536–548.
14. Kravitz DJ, Vinson LD, Baker CI (2008) How position dependent is visual object recognition? *Trends Cogn Sci* 12(3):114–122.
15. Ralph MAL, Jefferies E, Patterson K, Rogers TT (2016) The neural and computational bases of semantic cognition. *Nat Rev Neurosci*.
16. Peelen M V, Downing PE (2017) Category selectivity in human visual cortex: Beyond visual object recognition. *Neuropsychologia*.
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(3):314–25.
18. Wheatley T, Milleville SC, Martin A (2007) Understanding Animate Agents. *Psychol Sci* 18(6):469–474.
19. Weisberg J, Van Turennout M, Martin A (2006) A neural system for learning about object function. *Cereb Cortex* 17(3):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(2):432–444.
21. Kaiser D, Azzalini DC, Peelen M V (2016) Shape-independent object category responses revealed by MEG and fMRI decoding. *J Neurophysiol* 115(4):2246–2250.
22. Proklova D, Kaiser D, Peelen M V. (2016) Disentangling Representations of Object Shape

and Object Category in Human Visual Cortex: The Animate–Inanimate Distinction. *J Cogn Neurosci* 28(5):680–692.
23. Murty NAR, Pramod RT (2016) To What Extent Does Global Shape Influence Category Representation in the Brain? *J Neurosci* 36(15):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 M V, 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(1):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(6):687–692.
27. van den Hurk J, Van Baelen M, de Beeck HPO (2017) Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proc Natl Acad Sci*:201612862.
28. Bi Y, Wang X, Caramazza A (2016) Object Domain and Modality in the Ventral Visual Pathway. *Trends Cogn Sci* 20(4):282–290.
29. Wandell BA, Dumoulin SO, Brewer AA (2007) Visual Field Maps in Human Cortex. *Neuron* 56(2):366–383.
30. Golomb JD, Kanwisher N (2012) Higher level visual cortex represents retinotopic, not spatiotopic, object location. *Cereb Cortex* 22(12):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(3):479–490.
32. Larson AM, Loschky LC (2009) The contributions of central versus peripheral vision to scene gist recognition. *J Vis* 9(10):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(5):533.
34. Rajimehr R, Bilenko NY, Vanduffel W, Tootell RBH (2014) Retinotopy versus face selectivity in macaque visual cortex. *J Cogn Neurosci* 26(12):2691–700.
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(8):e1003167.
36. Nasr S, Echavarria CE, Tootell RBH (2014) Thinking Outside the Box: Rectilinear Shapes Selectively Activate Scene-Selective Cortex. *J Neurosci* 34(20).
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, de Beeck HO (2017) Avoiding illusory effects in representational similarity analysis: What (not) to do with the diagonal. *Neuroimage*.
39. Andrews TJ, Clarke A, Pell P, Hartley T (2010) Selectivity for low-level features of objects in the human ventral stream. *Neuroimage* 49(1):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(40):10111–10123.
41. Lerner Y, Harel M, Malach R (2004) Rapid completion effects in human high-order visual areas. *Neuroimage* 21(2):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(1):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(2):115–117.
45. Freeman J, Simoncelli EP (2011) Metamers of the ventral stream. *Nat Neurosci* 14(9):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(1). doi:10.1037/xge0000130.
47. Long B, Störmer VS, Alvarez GA (2017) Mid-level perceptual features contain early cues to animacy. *J Vis* 17(6).
48. Long B, Konkle T (2017) A familiar-size Stroop effect in the absence of basic-level recognition. *Cognition* 168.
49. Krizhevsky A, Sutskever I, Hinton G (2012) Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst*:1097–1105.
50. Srihasam K, Vincent JL, Livingstone MS (2014) Novel domain formation reveals proto-architecture in inferotemporal cortex. *Nat Neurosci* 17(12):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(4):e1003553.
53. Khaligh-Razavi SM, Kriegeskorte N (2014) Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Comput Biol* 10(11):e1003915.
54. Yamins DLK, et al. (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci U S A* 111(23):8619–24.
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(7):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(3):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(27):10005–10014.
58. Ponce CR, Hartmann TS, Livingstone MS (2017) End-stopping predicts curvature tuning along the ventral stream. *J Neurosci* 37(3):648–659.
59. Felleman DJ, Van Essen DC (1991) Distributed hierachical processing in the primate cerebral cortex. *Cereb Cortex* 1(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(39):12978–12995.
61. Kourtzi Z, Connor CE (2010) Neural Representations for Object Perception: Structure, Category, and Adaptive Coding. *Annu Rev Neurosci* 34(1):45–67.
62. Pasupathy A, Connor CE (2001) Shape representation in area V4: position-specific tuning for boundary conformation. *J Neurophysiol* 86(5):2505–2519.

63. Carlson ET, Rasquinha RJ, Zhang K, Connor CE (2011) A sparse object coding scheme in area V4. *Curr Biol* 21(4):288–293.
64. Yau JM, Pasupathy A, Brincat SL, Connor CE (2012) Curvature processing dynamics in macaque area V4. *Cereb Cortex* 23(1):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(1):313–319.
66. Rajimehr R, Bilenko NY, Vanduffel W, Tootell RBH (2014) Retinotopy versus Face Selectivity in Macaque Visual Cortex. *J Cogn Neurosci* 26(12):2691–2700.
67. 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(4):e1000608.
68. 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. doi:10.3389/fnhum.2016.00137.
69. Yue X, Pourladian IS, Tootell RBH, Ungerleider LG (2014) Curvature-processing network in macaque visual cortex. *Proc Natl Acad Sci* 111(33):E3467–E3475.
70. Konkle T, Oliva A (2011) Canonical visual size for real-world objects. *J Exp Psychol Hum Percept Perform* 37(1):23–37.
71. 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(4):676–697.
72. Konkle T, Caramazza A (2016) The Large-Scale Organization of Object-Responsive Cortex Is Reflected in Resting-State Network Architecture. *Cereb Cortex* 31(28):1–13.
73. Mahon BZ, Caramazza A (2011) What drives the organization of object knowledge in the brain? *Trends Cogn Sci* 15(3):97–103.
74. Willenbockel V, et al. (2010) Controlling low-level image properties: The SHINE toolbox. *Behav Res Methods* 42(3):671–684.

Supplementary Information for

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

Bria Long, Chen-Ping Yu, & Talia Konkle

Bria Long
Email: bria@stanford.edu

This PDF file includes:

5. Supplement to Predictive Modeling
    a. Voxel mask construction: **Figure S16**
    b. Modeling results in anterior to posterior regions: **Figure S17, Figure S18**
    c. Early visual cortex results and corresponding **Figure S19**

Other supplementary materials for this manuscript include the following:

All stimuli, pre-processed data, and main analysis code for this paper are available at the Open Science Repository for this project, https://osf.io/69pbd/, which is also linked to a GitHub codebase for generating texforms. Raw fMRI data is available on request.

1. Extended Methods

fMRI Data Preprocessing. Functional data were analyzed using Brain Voyager QX software and MATLAB. Preprocessing included slice scan-time correction, 3D motion correction, linear trend removal, temporal high-pass filtering (0.01 Hz cutoff), spatial smoothing (4 mm FWHM kernel), and transformation into Talairach (TAL) coordinates**.** Two subjects had one run in which they moved more than 0.5 mm within 2 seconds (1 TR) and these runs were discarded from analysis. The cortical surface of each subject from the high-resolution T1-weighted anatomical scan, acquired with a 3D MPRAGE protocol. To do so, we used the default segmentation procedures in FreeSurfer. Surfaces were then imported into Brain Voyager and inflated using BV surface module. Gray matter masks were defined in the volume and were constructed based on the Freesurfer segmentations.

General linear models (GLMs) were computed at the single subject level for texforms and original runs separately, both for the four main conditions (big animals, big objects, small animals, and small objects) as well as separately for the full set of nested conditions (each category x each classifiability level, 24 conditions total). GLMs included square-wave regressors for each condition's presentation times, convolved with a gamma function to approximate the hemodynamic response, fit to voxel-wise time course data with percent signal change normalization and correction for serial correlations. In Experiment 2, GLMs were fit eight main conditions of interest: each combination of category (big animals, big objects, small animals, and small objects) and visual field presentation (upper, lower) separately for texforms and originals.

MRI acquisition. Imaging data were collected using a 32-channel phased-array head coil with a 3T Siemens Prisma fMRI Scanner at the Harvard Center for Brain Sciences. High-resolution T1-weighted anatomical scans were acquired using a 3D MPRAGE protocol (176 sagittal slices; FoV = 256 mm; 1x1x1 mm voxel resolution; gap thickness = 0 mm; TR = 2530 ms; TE = 1.69 ms; flip angle = 7 degrees). For functional runs, blood oxygenation level-dependent (BOLD) contrast was obtained using a gradient echo-planar T2* sequence (84 oblique axial slices acquired at a 25° angle off of the anterior commissure-posterior commissure line; FoV = 204 mm; 1.5×1.5×1.5 mm voxel resolution; gap thickness = 0 mm, TR = 2000 ms; TE = 30 ms; flip angle = 80 degrees; multi-band acceleration factor = 3).

fMRI Experiment Design. Each run had twelve 6s blocks for each condition (big animals, big objects, small animals, small objects), with 10s rest periods interleaved every four blocks. Each block consisted of six images (5 unique images and 1 repeat) each presented for 800ms followed by a 200ms blank. Further, each block contained images from one of the six classifiability levels. Each classifiability level for each condition was shown twice per run. Thus, each texform image was shown twice during a run and 8 times over the entire experiment. All images were presented in isolation on a uniform gray background. This design choice allowed us to analyze neural responses in both a high-powered, four-condition design as well as a moderately powered 24-condition design for the predictive modeling analysis. Note, however, this does not allow us to model responses to individual texforms or their corresponding recognizable images.

In Experiment 1, each image subtended 10.36° x 10.36° visual angle centered at fixation. In Experiment 2, each block of images could appear either above or below fixation (6.92° x 6.92° degrees of visual angle, bottom edge .86° degrees from center). These positions were counterbalanced across blocks such that, for each level of classifiability and condition, one block was presented in the upper visual field and the other block was presented in the lower visual field. Participants were instructed that maintaining fixation was more important than task performance, and fixation was monitored online using an EyeLink

1000 eye-tracker. Participants were calibrated to the eye-tracker at the beginning of the experiment and were recalibrated every 2-3 runs as needed. See **Figure S12** for fixation heatmaps for each participant.

Retinotopy Protocol. Additionally, participants completed a retinotopy protocol in order to define early visual areas V1-V3. Observers viewed bands of flickering checkerboards in a blocked design. The conditions included vertical meridian bands (~22° × 1.7°), horizontal meridian bands (~22° × 1.7°), iso-eccentricity bands covered by a central ring (radius ~1.2° to 2.4°), a peripheral ring (radius ~5.7° to 9.3°), and an extra wide peripheral ring (inner radius ~9.3°, filling the extent of the screen). In Experiment 2, the vertical and horizontal meridian bands were replaced with wedges. The apex of each wedge was at fixation and the base extended to ~22° in the periphery, and the checkerboard patterns flickered at 6 Hz. Each block was 6 seconds, within which the checkerboard cycled at 8 Hz between states of black-and-white, randomly colored, white-and-black, and random colored. In each 4.4-min run (142 volumes), the 5 visual field band conditions and 1 fixation condition were repeated 7 times with their order randomly permuted within each repetition. Each run started and ended with a 6 s fixation period. Participants' task was to maintain fixation, and press a button every time the fixation dot turned red, which happened once per block. Using data from this retinotopy protocol, early visual regions (V1-V3) were defined by hand on inflated brain guided by the contrast of horizontal vs. vertical meridians (see (1)).

Preference Map Correlation Details. Two comparisons were used to assess map-correlation robustness. First, we compared map correlations to a shuffled voxel baseline. For each subject, the spatial position of texform voxels was shuffled and then correlated with the unshuffled original preference map. This was repeated 1000 times, yielding a chance distribution for each subject, from which a p-value was computed based on how often the simulated shuffled values were greater than the observed map correlation. Second, we considered the map correlations relative to an estimated noise ceiling. To do so, in each subject, texform preference maps were correlated between odd and even runs, yielding a texform map split-half correlation. The same analysis was repeated for originals. If any of these odd-even correlations was less than zero (i.e., a negative correlation), we substituted this value with zero; this occurred in one subject for the object size comparison. Given these split-half texform and original map correlations were estimated with half the power of the texform-original map comparison, we used the Spearman Brown prophecy formula to approximately adjust the reliabilities (N*observed reliability / 1 +(N-1)*observed reliability, where $N = 2$ as we divided the data in half). Then, the noise ceiling for the texform-original map correlation was computed separately for each participant, as the square root of the product of these corrected reliabilities.

Posterior-to-Anterior Correlations. To assess whether there was a difference in the overall strength of the original animacy preferences vs. the texform animacy preferences along the posterior-to-anterior gradient, we computed difference scores for each anatomical section for each participant. We then performed a simple rank correlation between ascending anatomical sections (i.e., 1,2,3,4,5) and these difference scores (originals – texforms) in each subject. A rank correlation metric was used to assess whether originals generate greater animacy preferences along this posterior to anterior gradient, without assuming a meaningful relationship with the TAL-Y coordinates of the anatomical sections. Finally, we asked whether these rank correlations were above zero at the group level by performing a one-sample t-test over subjects. The same procedure was repeated for the object size distinction. For visualization purposes, in **Figure S9A-C**, we also defined group-level anatomical sections based on the Group GLM activations, with the scatter plots showing voxel response preferences for animacy and size dimensions based on the group GLM beta fits.

Predictive Modeling Feature Spaces.

Gabor & Gist Models. Gabor features were extracted in an 8 x 8 grid over the original, recognizable images (440 x 440 pixels) at three different 3 scales, with 8,6, and 4 oriented Gist per scale, respectively

(Oliva & Torralba, 2006). GIST model features were extracted by taking the first 20 principle components of this Gabor feature matrix. In both cases, these features were then averaged across the five images in each nested classifiability group presented during the fMRI experiment. The squared Euclidean distance along each feature was used to construct feature RDMs for use in the predictive modeling procedure, and all dissimilarities were scaled between 0 and 1, yielding a 276 x 896 feature vector for Gabor features and a 276 x 20 feature vector for Gist features.

Texture Synthesis Model (Freeman & Simoncelli, 2011): The texture synthesis model has 10 feature classes (corresponding to pixel statistics, weighted marginal statistics, simple cell responses, complex cell responses, cross-position correlations (i.e., autocorrelation) within scales computed separately for simple and complex cells, cross-orientation correlations computed separately for simple and complex cells, and cross-scale correlations computed separately for simple and complex cells). Features were included if they (1) had any variance across the images (SD > 0) and (2) were calculated within pooling windows tiling the depicted item (see **Figure S1** for an illustration of the pooling windows). The values for each feature were then z-scored across the 120 images, and then averaged over the five images in each classifiability group. Each feature was converted to an RDM using squared Euclidean distance, and all dissimilarities were scaled between 0 and 1, generating a 276 x 20,914 feature matrix.

Behavioral Ratings–Animacy/Size: For texforms, feature RDMs were constructed based on the behavioral animacy and size classifiability scores. Note these are the same scores used to group the texforms into the nested design. These experiments yielded a vector corresponding to participants ability to classify each texform as an animal (range: 0-1, where 1 = always classified as an animal, and 0 = never classified as an animal) and their ability to classify each texform according to their size in the real world (range: 0-1, where 1 = always classified as big in the real-world, and 0 = never classified as big in the real world). These scores were averaged according to the 24 nested conditions presented during the experiment, yielding a 24 x 1 vector for animacy and a 24 x 1 vector for size for texforms and for originals. We then took the squared Euclidean distance of each 1-dimensional feature vector and all dissimilarities were scaled between 0 and 1; the final feature matrix was a 276 x 2 feature matrix.

For original images, feature RDMs were constructed using their actual animacy/size in real-world. These yielded a vector corresponding to the actual animacy of the recognizable image (1 = animate, and 0 = inanimate), and a vector corresponding the actual size of the object in the real world (1 = big in the real-world, and 0 = small in the real world). We then took the squared Euclidean distance of each 1-dimensional feature vector and all dissimilarities were scaled between 0 and 1; the final feature matrix was a 276 x 2 feature matrix.

Behavioral Ratings–Curvature: Behavioral ratings on Amazon Mechanical Turk were obtained to assess the perceived curvature of both the texforms and the originals; 30 participants rated the curvature of the 120 texforms, and another 30 participants rated the curvature of their 120 corresponding original images. Participants were asked, "How boxy or curvy is the thing depicted in this image?" and asked to respond using a 1-5 scale (1: Very curvy, 2: Mostly curvy, 3: Equally boxy and curvy, 4: Mostly boxy, 5: Very boxy). See **Figure S4A** for an illustration of the task. These ratings were averaged across participants, and then averaged across the five images in each classification group. This yielded two 24 x 1 vectors corresponding to the average perceived curvature of each group of texforms and of each group of original images. The squared Euclidean distance of all pairwise comparisons of these conditions was computed separately for texforms and originals, yielding two 276 x 1 feature vectors for curvature for modeling responses to texforms and originals; all dissimilarities were scaled between 0 and 1. See **Figure S4B** for a visualization of this data and a comparison of the curvature ratings between texforms and recognizable images.

CNN Features (Texforms & Originals): The AlexNet architecture (2) as was trained using the conventional image classification task using the ImageNet dataset. The standard AlexNet training regime was adopted using a public code package (https://github.com/soumith/imagenet-multiGPU.torch) that was optimized for multi-threaded CNN training in Torch7. Specifically, stochastic gradient descent (SGD) optimization was used with 0.9 momentum, an initial learning rate of 0.02, and weight decay of 0.0005. Both the learning rate and the weight decay follow a pre-defined decreasing schedule (see train.lua from the code package) using a mini-batch size of 128, with 10,000 mini-batches per epoch over a total of 55 training epochs. Standard data augmentation such as random horizontal flips and random 224-by-224 crops were performed during training.

Using the fully-trained network, CNN features were extracted from each unit in the CNN from both original and texform image sets. Specifically, for each image and each convolutional filter, we computed the summed activation map of the filter (an m-by-m map where m is the output size of the convolutional layer), accounting for border effects by setting to zero all values in the activation map within 10% of the four edges. This procedure yielded five feature matrices for the original images of 120-by-64, 120-by-192, 120-by-384, 120-by-256, and 120-by-256, corresponding to each of the five convolutional layers, and another set corresponding to the texform images. For the two fully-connected layers (layer 6 & 7), the activation level to each image was direct computed (no global summation required), resulting in two feature matrices of 120-by-4096, corresponding to layer 6 and 7, and another set for texform images.

Each feature matrix was normalized by dividing the rows with its L2-norm, and the rows were averaged over the five images in each classifiability group. Finally, for each feature (column) of each feature matrix, we computed the pairwise squared Euclidean distance between all of the 24 conditions, yielding five 276-by-m representational dissimilarity matrices, where m is the number of convolutional filters for the corresponding layer. All dissimilarities were then scaled between 0 and 1. These RDMs were used to perform feature modeling of the individual CNN layers.

2. Extended Materials: Texform stimuli details

Here, we provide additional details on (a) how the stimuli set was constructed (**Fig. S1**) (b) an overview of how basic-level recognition was assessed (**Fig. S2A**), (c) how classifiable the texforms are by their animacy/size, how these were used to form the nested groupings, and relationship between basic-level recognition and classifiability (**Fig. S3, Fig. S2B**), (d) the perceived curvature of the texforms and recognizable images (**Fig. S4)** and (e) how recognizable the texforms were after the neuroimaging session (**Fig. S5**).

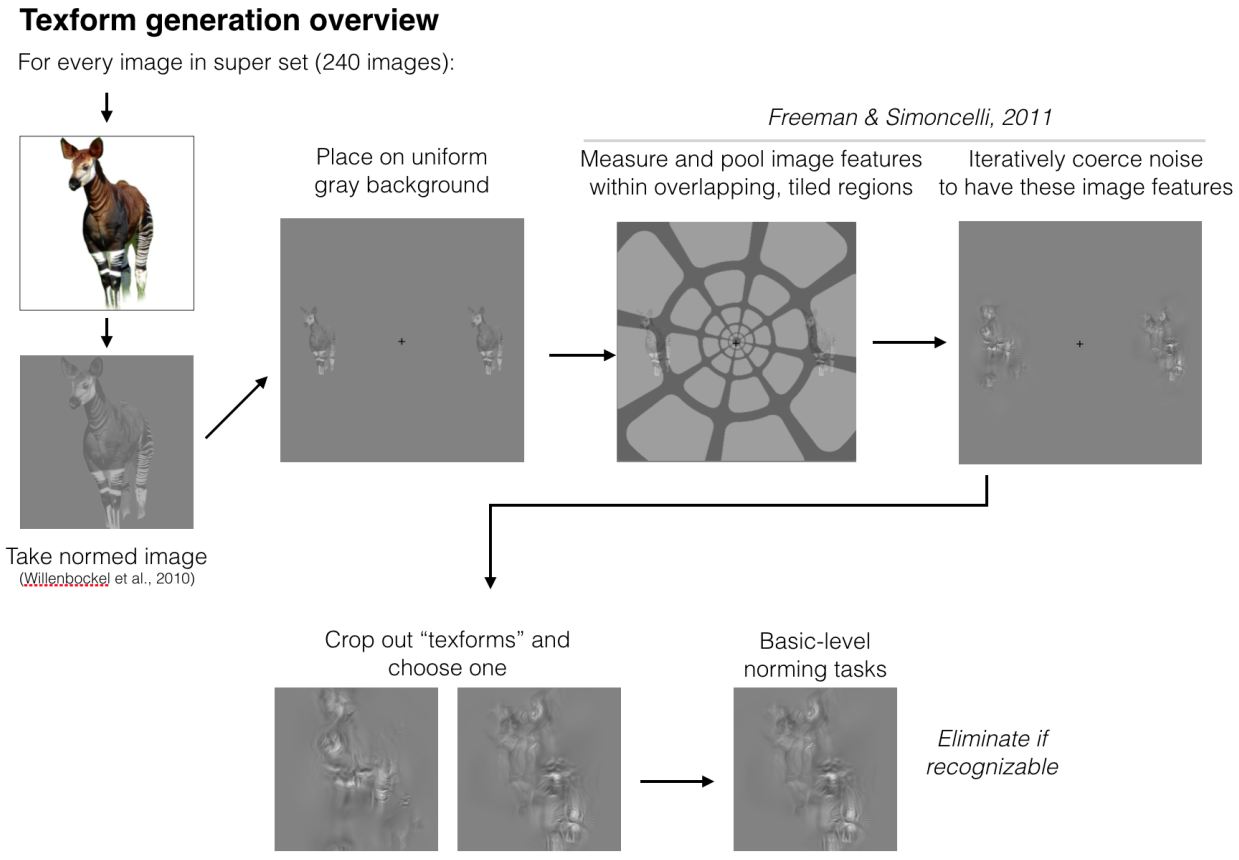

Fig. S1. Overview of the procedure used to generate texforms. Original, color, images were converted to grayscale and matched for overall luminance and contrast within the superset of 240 images. Next, these normed images were placed in the "periphery" of the model on the uniform gray background. First and second-order image statistics (3) were measured and pooled within overlapping, tiled regions illustrated here (pooling window parameters, scaling = .5, AR = 1). Next, synthetic stimuli were generated by coercing , random noise to have the similar image statistics within these pooling windows. The procedure was run for 50 iterations using a variant of gradient descent. This produced an synthesized image with two texforms, which were then cropped out. Norming tasks were then used to select a set of texforms that were unrecognizable at the basic-level (see below).

Texform Selection Details. Online recognition experiments were run to assess how recognizable each texform was. First, 18 participants guessed the identity of each of 240 texform images. Next, six new participants assessed the validity of these guesses. These participants were presented with the original images and all of the texform guesses, and judged whether each guess could be "used to correctly describe" the original image (see **Figure S2A**). The proportion of guesses accepted as correct yielded a basic-level identification score for each image. Images in which a rater accepted more than 3/18 responses as correct were removed. Next, 120 texforms and their corresponding originals were selected (30 images per category), with the constraints that the categories did not significantly differ in either aspect ratio or pixel area; all $p >= .1$). On average, these 120 texforms were identified at the basic level <3% of the time. Finally, the overall luminance and contrast levels across all 240 images (120 texforms, 120 originals) equated using the SHINE toolbox (4) and the edges of all of images were blurred so that they gradually faded into their backgrounds.

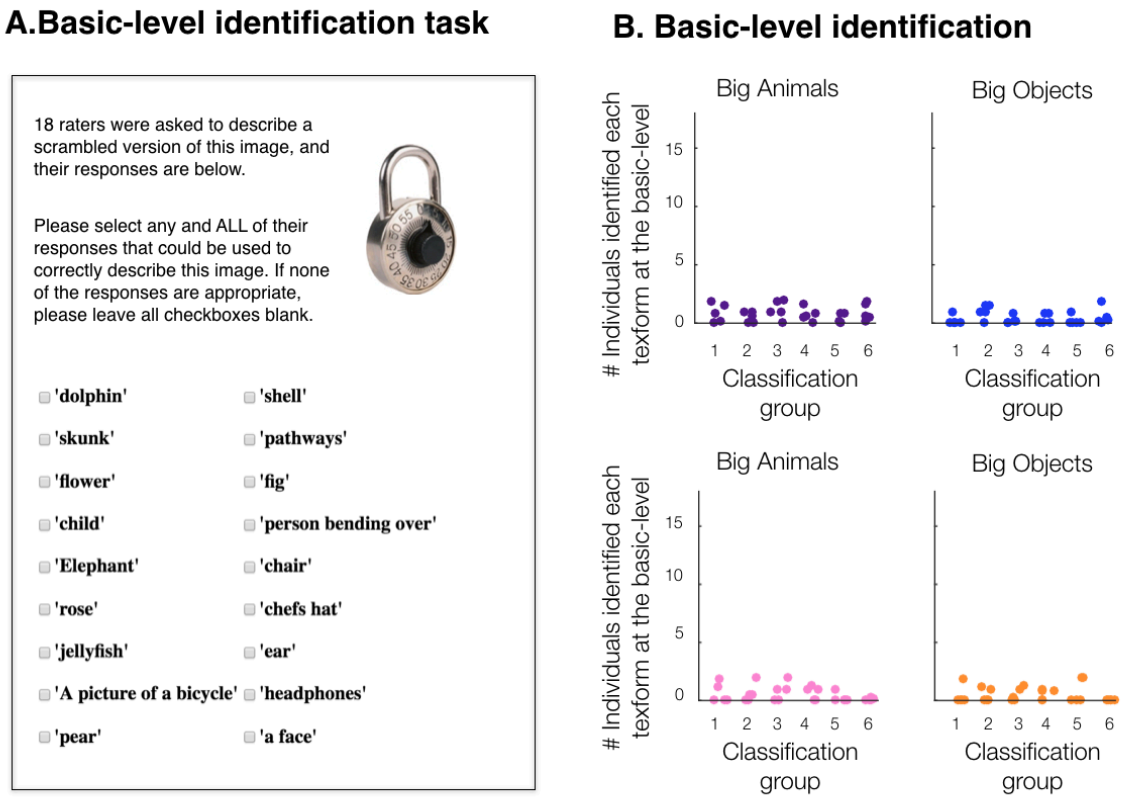

Figure S2. (A). Example trial from the basic-level identification task; raters determined whether the guesses could be used the describe the original image from which the texform was generated. (B). Basic-level identification rate from the 18 norming participants as a function of animacy/size classification group for each object category, each point represents a texform image. X-axis position is jittered to show all points.

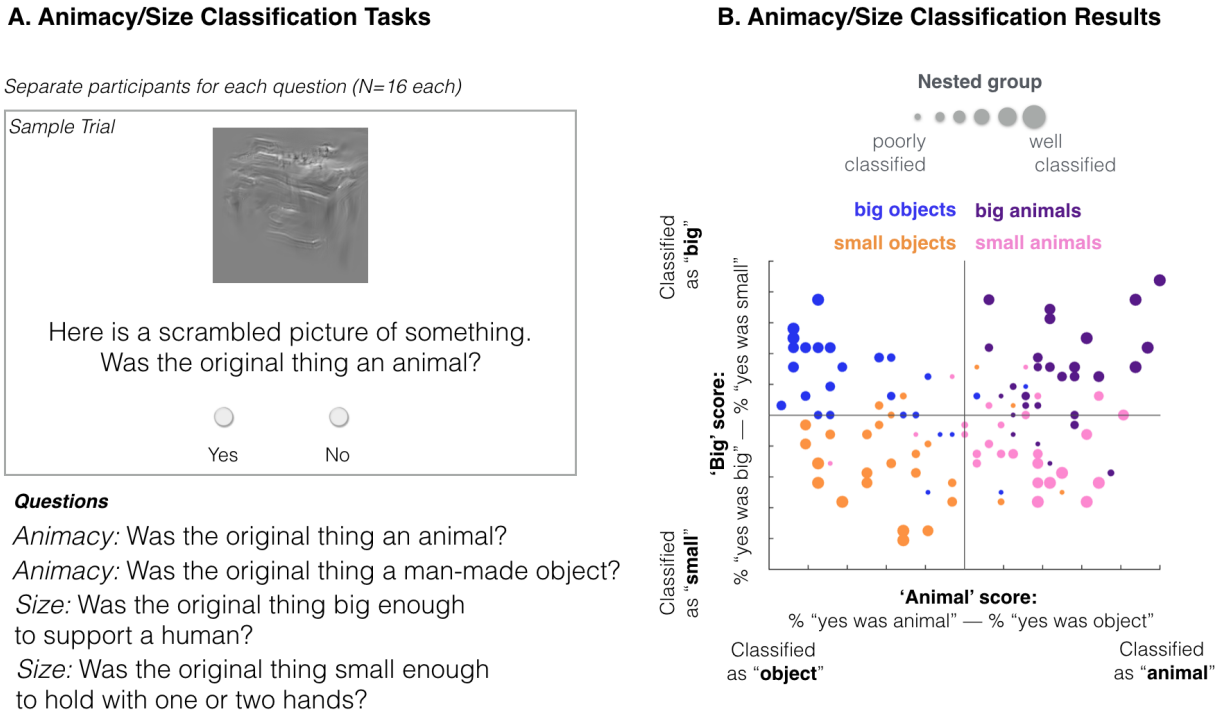

Figure S3. (A) Schematic of the animacy/size classification tasks. (B) The classifiability each image plotted for animacy (x-axis) and real-world size (y-axis); each dot corresponds to a texform image. The position of the dot reflects its classifiability score on both axes, the color of the dot indicates the actual condition of the texform (big/small animal/object), and the size of the dot indicates which of the 6 classifiability groups it was assigned to, where larger dots represent groups of texform images that were better classified by their animacy and real-world size.

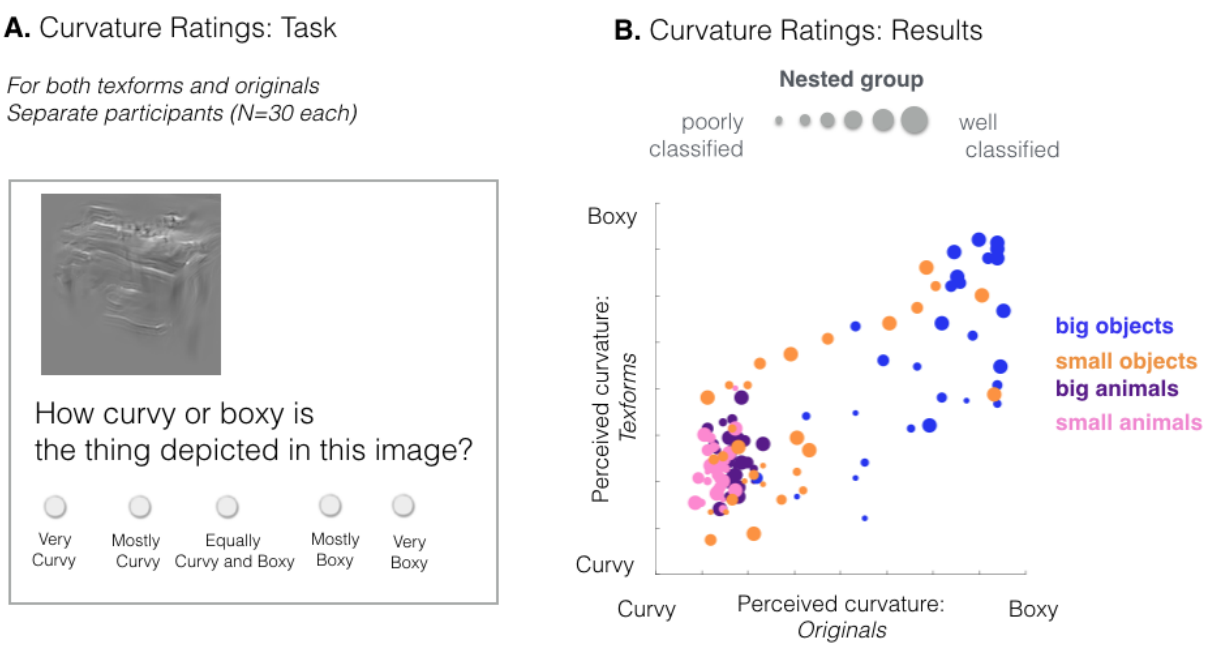

Figure S4. (A) Schematic of the perceived curvature task. (B). The perceived curvature of the texforms (y-axis) is plotted as a function of the perceived curvature of the recognizable, original images (x-axis); larger dots represent groups of images that were better classified by their animacy and real-world size.

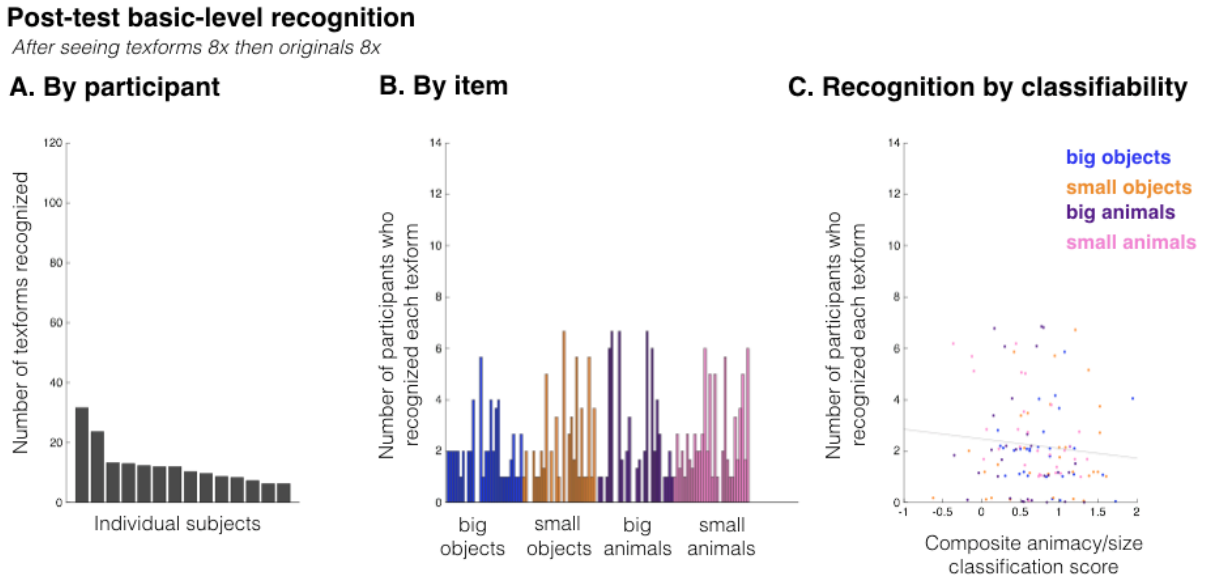

Figure S5. *Post-test* recognition results for the neuroimaging participants. After the scanning session, participants were told that the texforms actually were generated from real-world objects, and then completed a task in which they guessed what each texform might be. Note that all observers had seen each texform image eight times, then each original image 8 times, while in the scanner, before this test was taken. Three naïve observers rated whether the participants' texform guesses could be used to describe the original images, and they were told to be generous with what they counted as correct. (A) The number of texforms recognized by each participant is plotted. (B) The number of participants who recognized each texform is plotted; texforms are ordered according to their condition and composite classifiability score. (C) The item effects in B are plotted by the classifiability score of each item. We did not find strong evidence that the more classifiable texforms were also the ones that were more likely to be recognized after the scanning session; if anything, the trend was in the opposite direction.

3. Supplement to Experiment 1

Here, we first (a) provide additional details on how active OTC voxels were selected (**Fig. S6**), (b) show group preference maps (**Fig. S7**) and all individual subject preference maps (**Fig. S8**) for the animacy and object size distinctions, (c) plot group tripartite maps for direct comparison with Konkle & Caramazza (2013) (5) (**Fig. S9),** (d) plot posterior-to-anterior scatterplots of group-level animacy and object size preferences (**Fig. S10**), and (e) illustrate overall differences in response magnitude to texforms vs. recognizable images (**Fig. S11**).

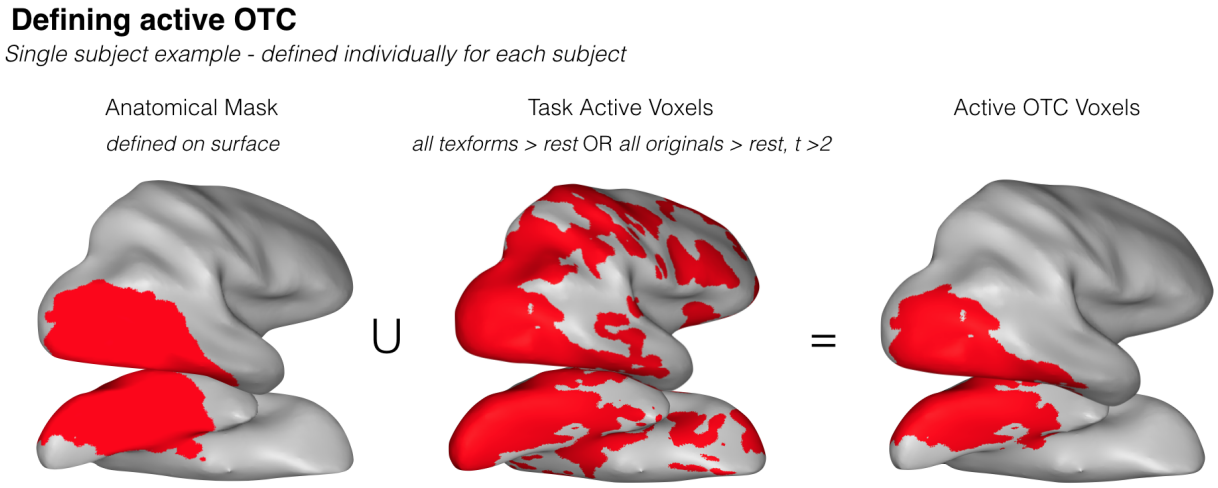

Figure S6. Schematic of how active OTC voxels were defined for use in Experiment 1. An anatomical mask of the occipitotemporal cortex was defined on the surface, with early visual regions (V1-V3) localized from the retinotopy protocol removed (left panel). Task-active voxels were defined from the contrast of all conditions > rest with $t > 2$ in either texform runs or original runs (middle panel). Active OTC was taken as the intersection of these two masks and was used for subsequent analyses. This procedure was carried out in each participant.

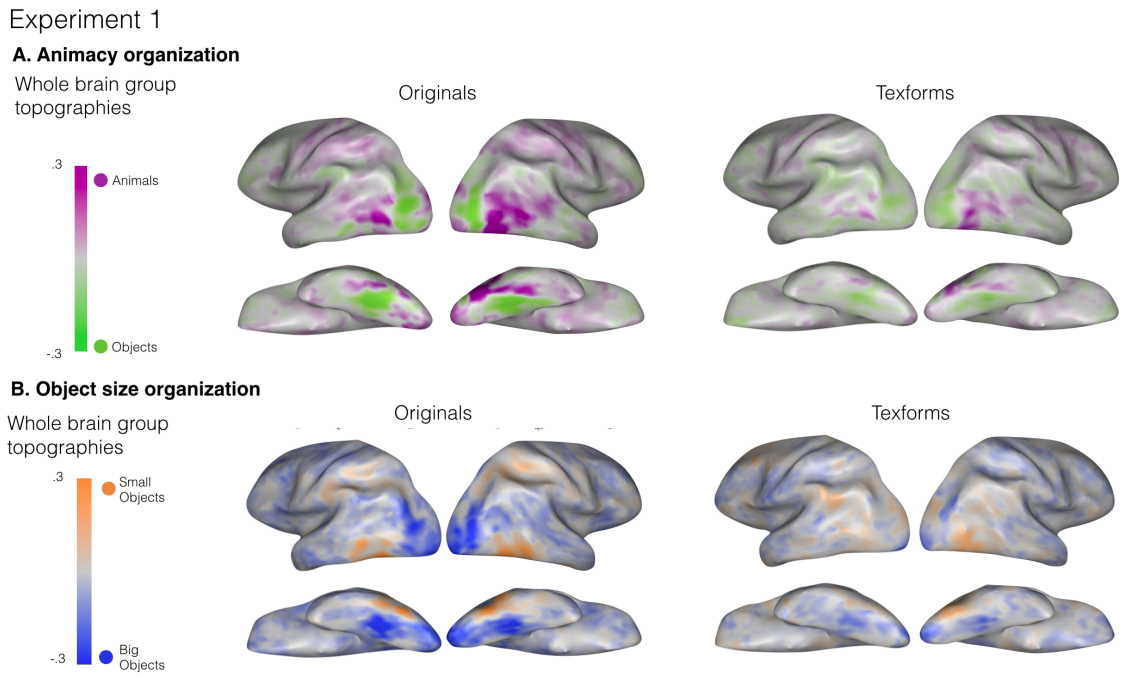

Figure S7. Whole brain group topographies for the animacy and object size distinctions, shown separately for originals (left panels) and texforms (right panels).

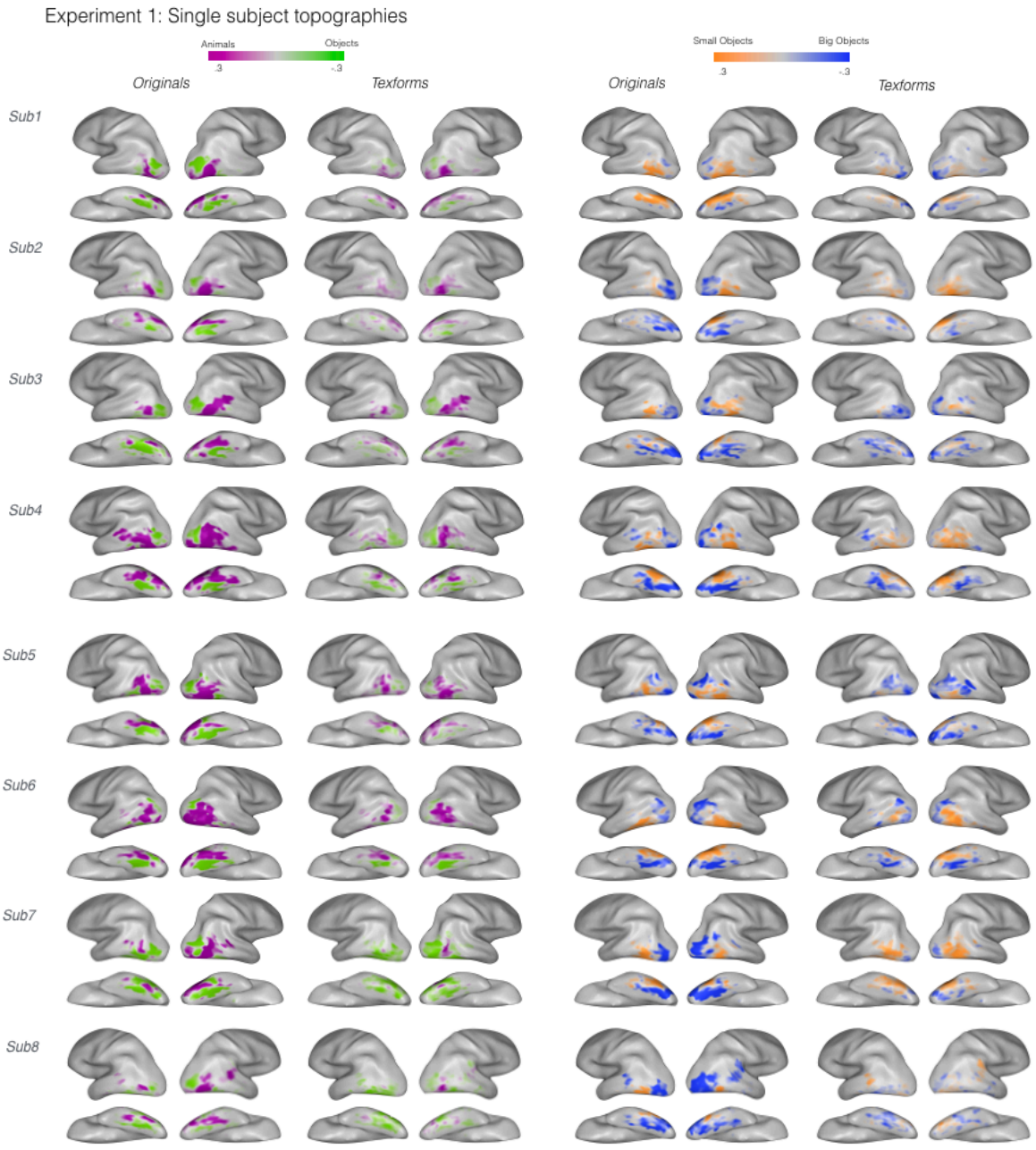

Figure S8. All single-subject topographies in both hemispheres for the animacy (left panels) and object size (right panel) distinctions, shown separately for originals and texforms. Preferences are shown within task-active occipito-temporal voxels.

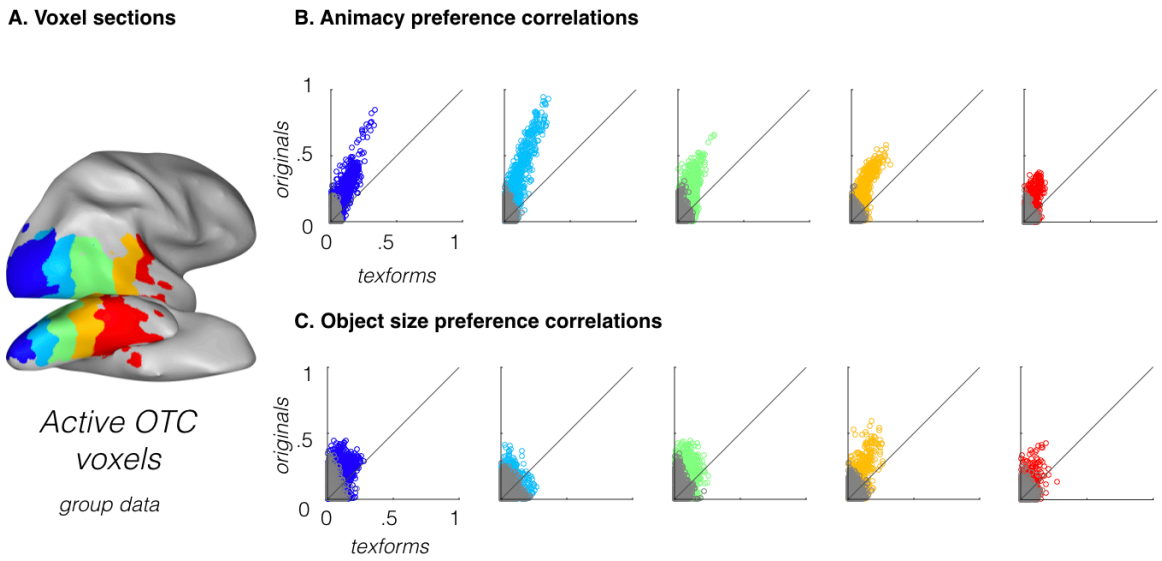

Figure S9. (A). Anatomical sections (shown here at the group level) from posterior to anterior in blue to red. (B) The animacy preferences elicited by texforms (x-axis) and by originals (y-axis) are plotted for each of the anatomical sections in the five subplots. Each point is a voxel. The x- and y-axes show strength of the animacy preference, computed as the absolute value of the difference between animal and object beta values. All points above the diagonal are voxels that show stronger animal/object preferences for original images than for texforms. Voxels where texforms and originals did not show the same preference are plotted in grey. (C) Object size preferences elicited by texforms (x-axis), and originals (y-axis) for each anatomical section.

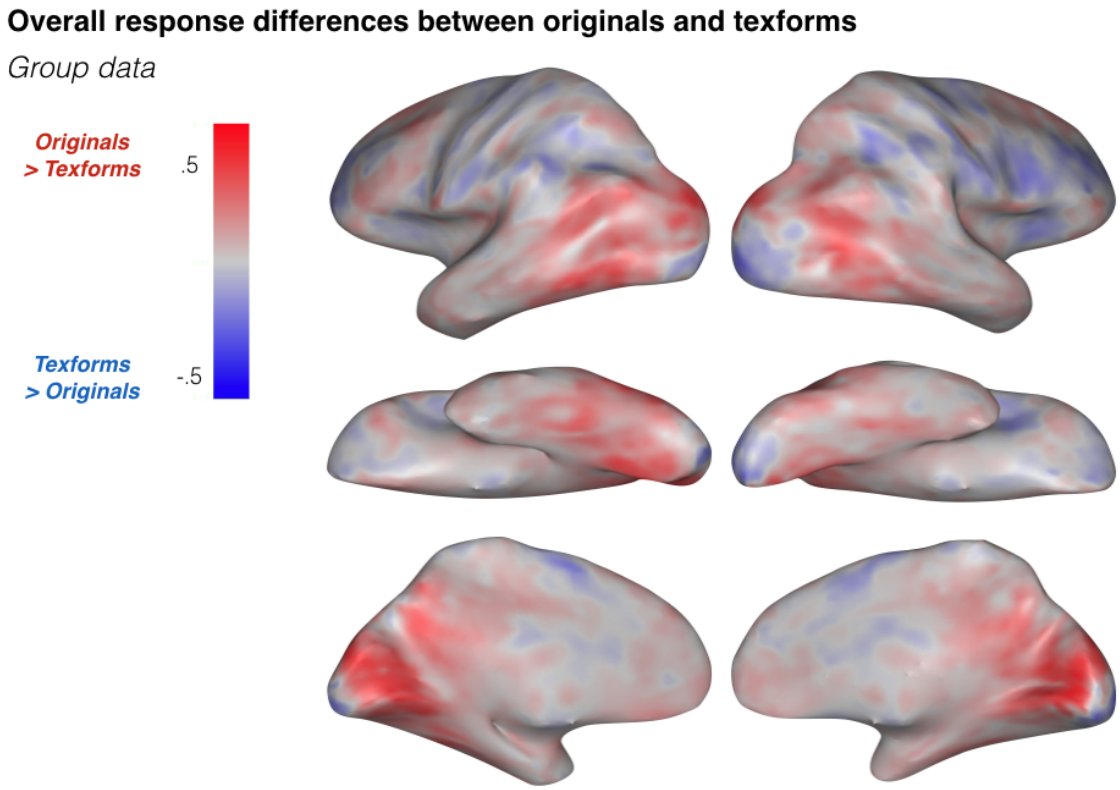

Figure S10. Overall activation differences between originals and texforms are shown at the group level. Voxels that showed stronger responses to originals are colored in red, and voxels that showed greater response differences to texforms are colored in blue.

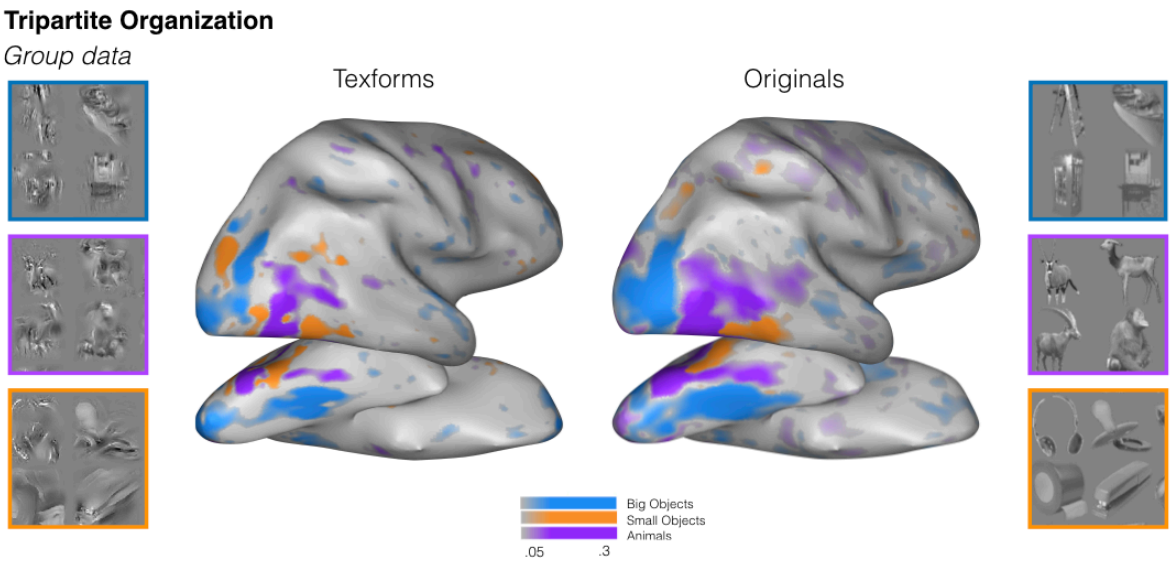

Figure S11. Comparison with Konkle & Caramazza, 2013 (5). (A) Group tripartite preference maps shown for both texforms (left) and originals (right) within task-active voxels.

4. Supplement to Experiment 2

Below, we show (a) maps of fixation stability for each participant (**Fig. S12**), (b) our procedure for defining location-tolerant voxels (**Fig. S13**), and (c) both group conjunction preference maps (**Fig. S14**) and all individual conjunction preference maps (**Fig. S15).**

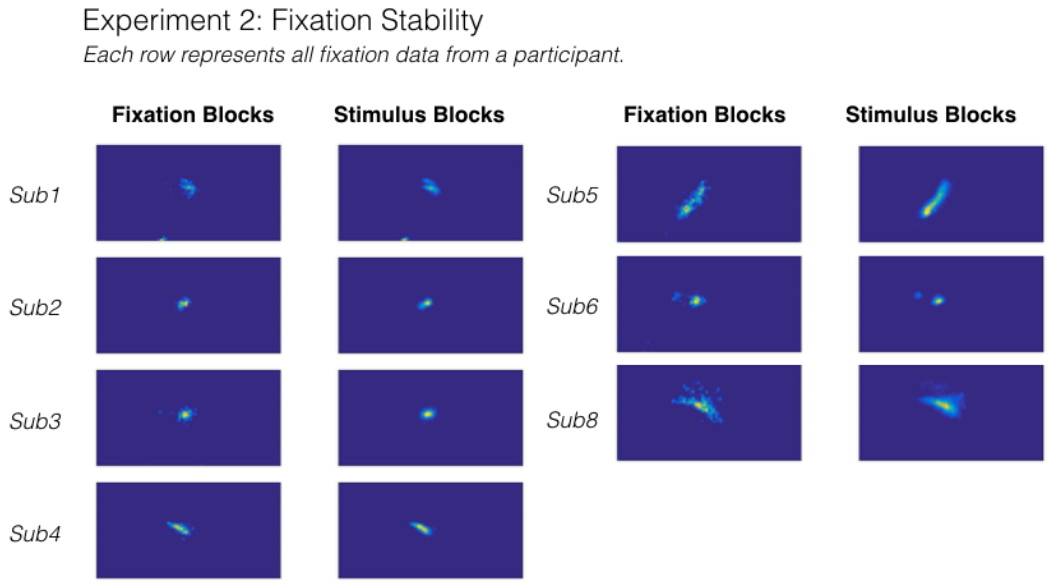

Figure S12. Fixation distributions are shown for each of the seven participants of Experiment 2 for whom we have eye-tracking data. The left column shows fixation distributions during time periods in which only a fixation dot was on the screen, and the right panel during time periods in which the stimuli were on screen at either an upper or lower visual location (in addition to the fixation dot). Note that while we were unable to obtain accurate calibrations for each participant, the deviations from fixation are highly similar between fixation and stimulus blocks. Thus, it is likely that this deviations from tight fixation reflect drift/noise in the calibration, rather than systematic looks towards the upper or lower visual field.

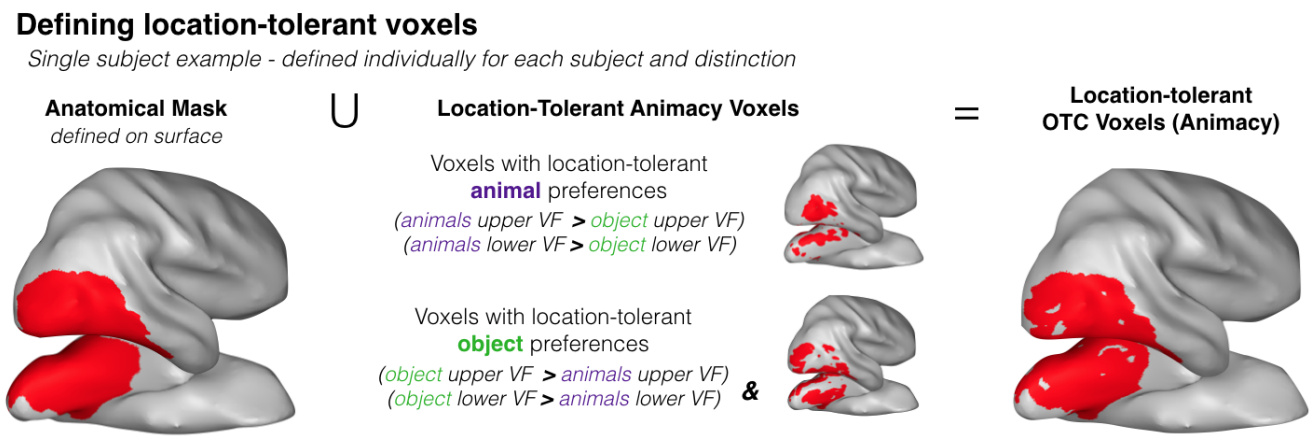

Figure S13. The procedure for defining location-tolerant voxels is shown for a single participant. First, an anatomical mask was defined on the surface for each participant to include occipitotemporal voxels and exclude early visual voxels (V1-V3) localized from a separate retinotopy protocol. Next, location-tolerant voxels were computed for each contrast (e.g., animals > objects and objects > animal) within this anatomical mask (middle panel). Finally, these two sets of location-tolerant voxels that prefer animals and objects, respectively, were fused together to create the final conjunction mask. The same procedure was followed for the object size distinction. These masks were computed separately for original images and texform images, in each participant.

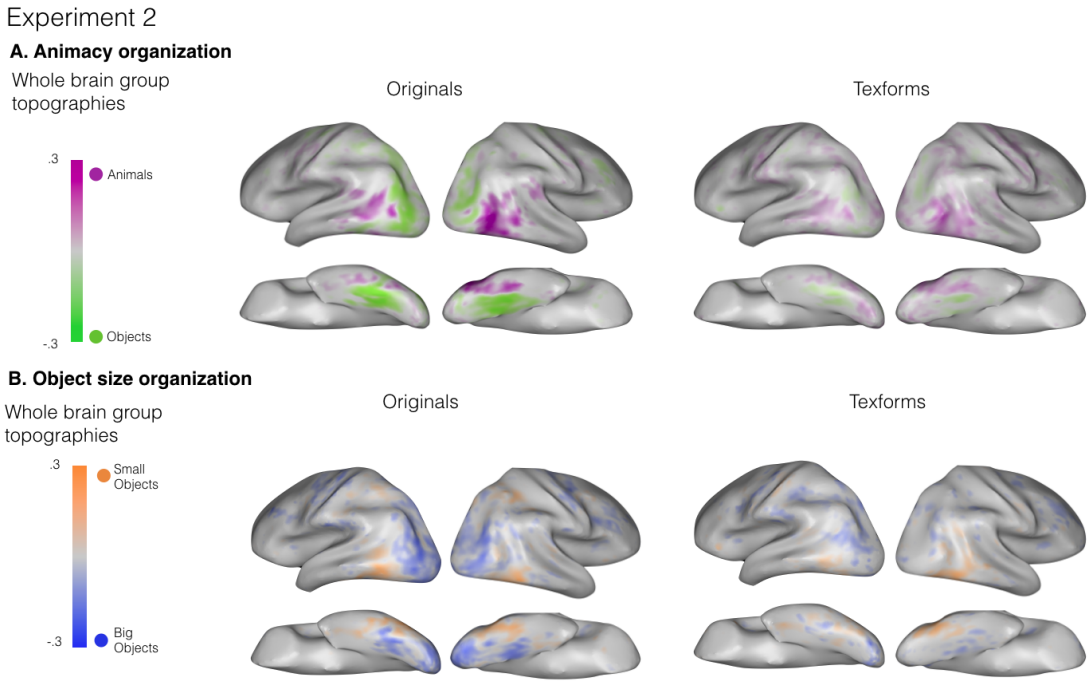

Figure S14. Group-level conjunction topographies for animacy (top) and object size (bottom) for texforms and originals. Preferences are shown within location-tolerant OTC voxels to originals (see Fig. S13).

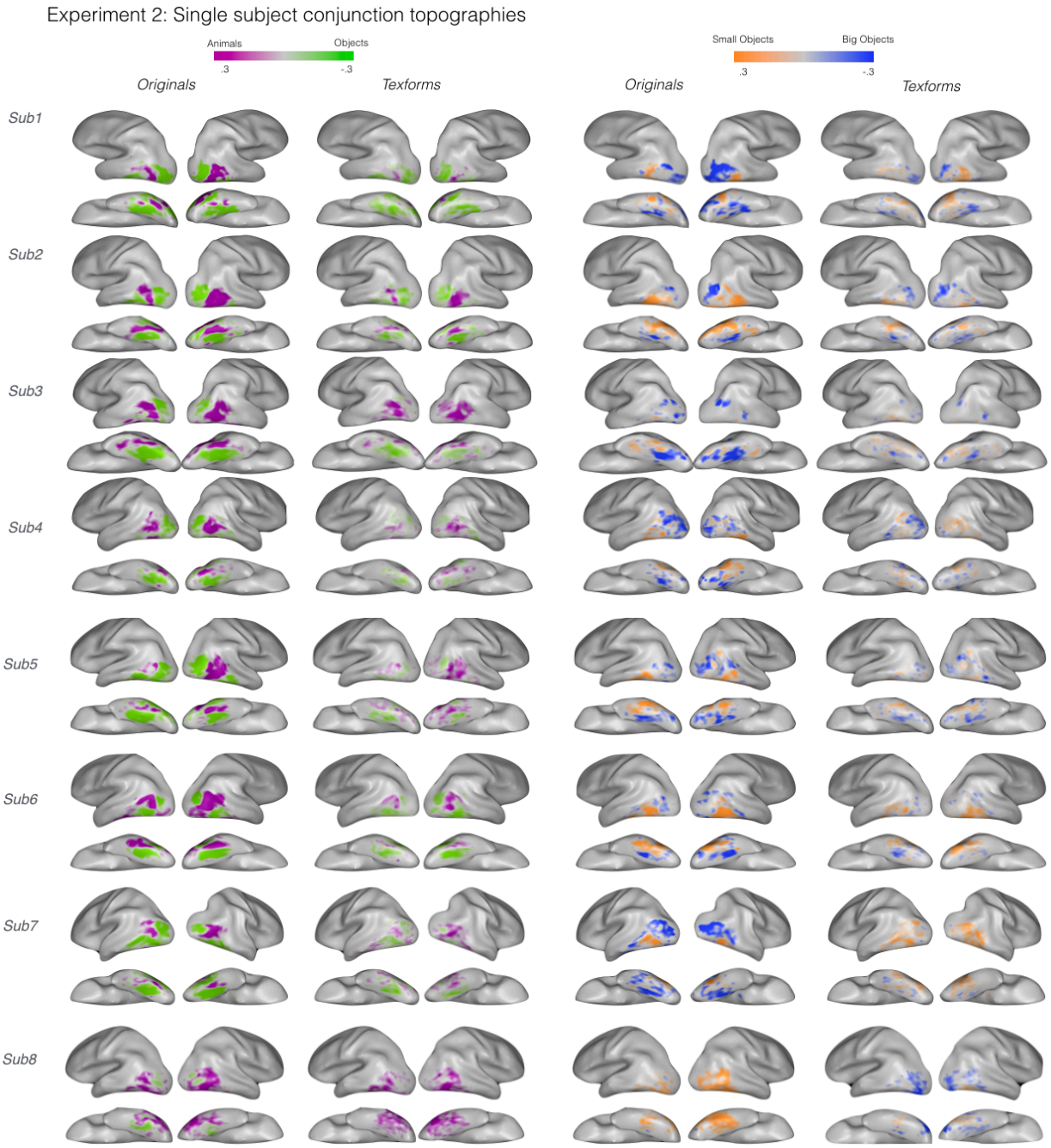

Figure S15. All single-subject conjunction topographies are shown in both hemispheres for the animacy (upper panel) and object size (lower panel) distinctions, shown separately for originals and texforms. Preferences are shown within location-tolerant OTC voxels to originals (see Fig. S13).

5. Supplement to Predictive Modeling Analyses

Below, we illustrate how reliable, task-active voxels were selected for the predictive modeling analyses (**Fig. S16**). We then report the results of additional modeling analyses long a posterior – anterior gradient of the ventral stream (**Fig. S17, S18)** and in early visual cortex (**Fig. S19**).

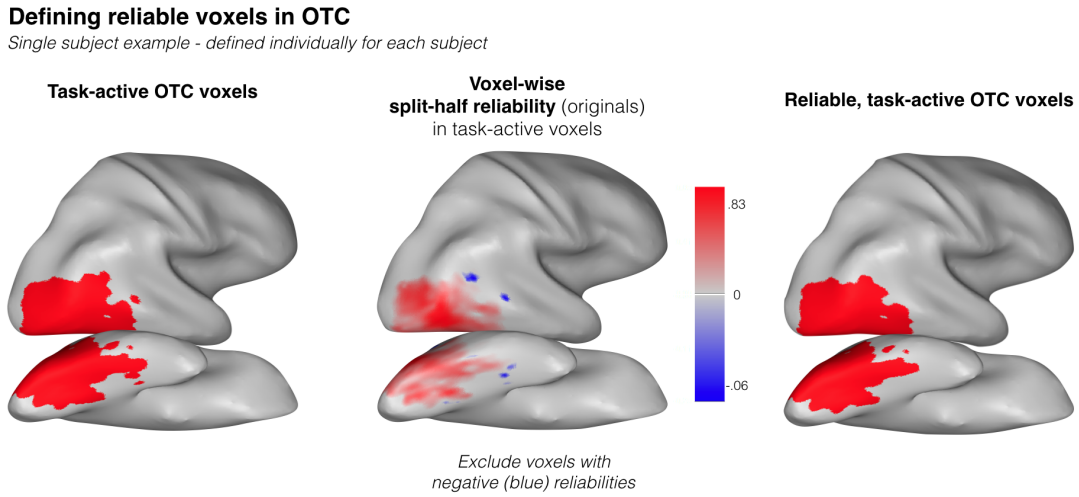

Figure S16. The procedure for reliable OTC voxels is shown for a single participant. First, we started with the task active voxel mask, as defined in **Figure S6**. Next, for each of these voxels, we extracted data from the condition-rich design (24 conditions), in odd and even runs in which original images were presented. The correlation between these two activation profiles was computed, and these voxel-wise split-half reliabilities are plotted in the middle panel. Voxels that are colored blue have slightly negative reliabilities. For the predictive modeling analysis, we excluded any voxel with a split-half reliability below zero. The right panel shows the final set of selected voxels for analysis for this participant. This voxel selection procedure was made a priori, with the motivation that it makes sense to only allow voxels that positively correlate with themselves in odd-even halves of the data to contribute to the final neural RDM. That being said, we also tested the impact of this choice in a post-hoc analysis, by repeating the analyses on the full set of active OTC voxels. The modeling results were extremely similar both qualitatively and quantitively.

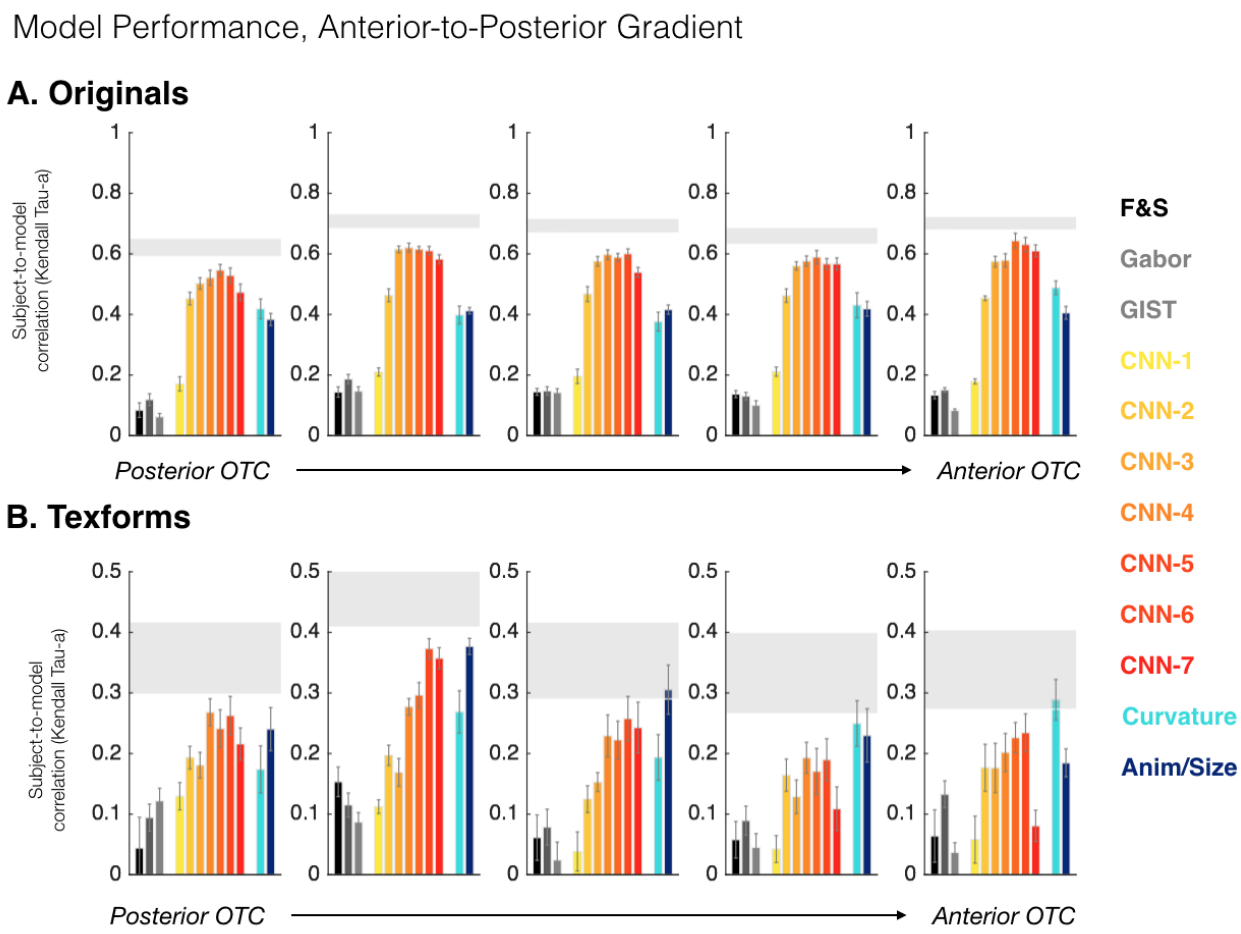

Figure S17. Predictive modeling results by anatomical sections of occipito-temporal cortex for originals (A) and texforms (B).

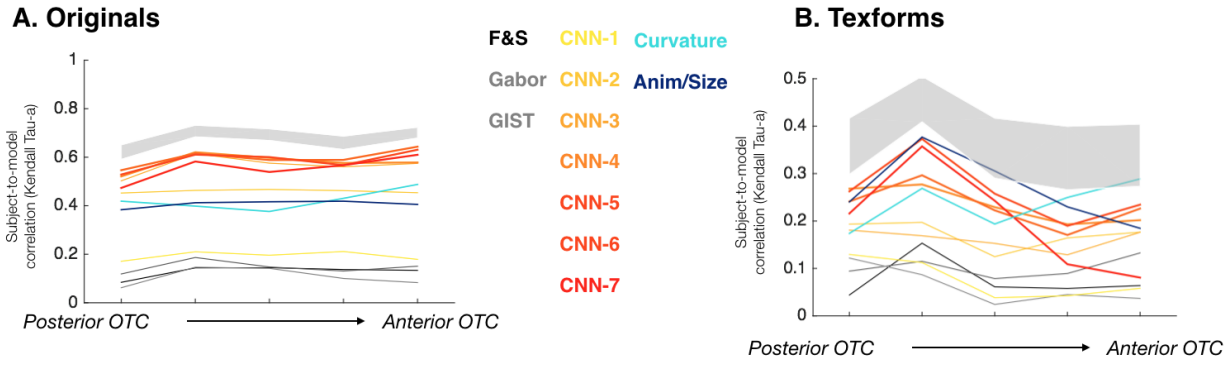

Figure S18. A summary of the results from **Fig. S17** are shown, where the performance of each model across regions is plotted for both originals (A) and texforms (B).

Modeling in Early Visual Cortex (EVC). Given that we generated texforms using a texture synthesis model that explains variance in V2/V4 (6, 7), we explored which feature spaces explained variance in early visual cortex by applying the same analytic method. Consistent with prior work, we found that Gabor-based models explained the most variance in early visual cortex, whereas models based on perceptual properties (e.g., perceived curvature) or category-based models explained less variance.

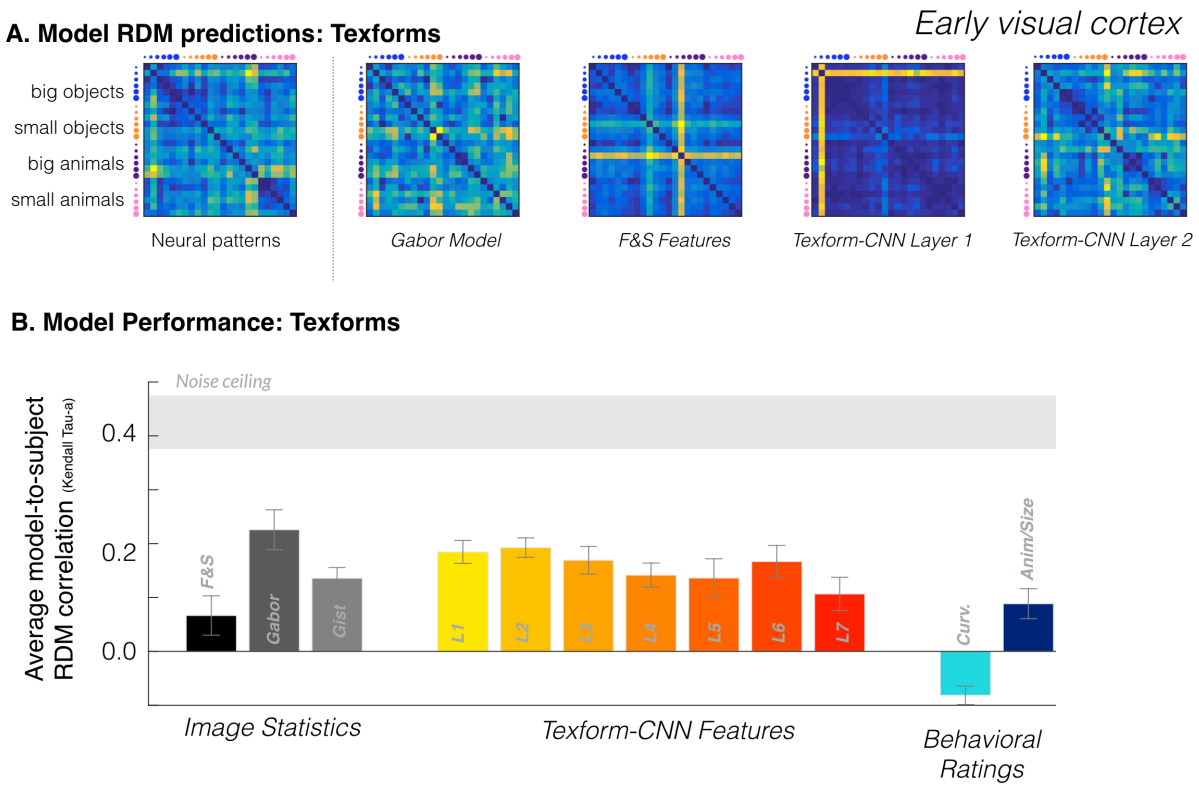

Figure S19. (A) Neural patterns in response to texforms in *early visual cortex* and predicted neural dissimilarities for selected models obtained through the same cross-validation procedure. (B) Average predicted model correlation (Kendall Tau-α) with individual subjects' neural patterns in EVC. Data is plotted with respect to the noise ceiling, shown in light gray.

References:

1. Cohen MA, Konkle T, Rhee JY, Nakayama K, Alvarez GA (2014) Processing multiple visual objects is limited by overlap in neural channels. *Proc Natl Acad Sci* 111(24):8955–8960.
2. Krizhevsky A, Sutskever I, Hinton G (2012) Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst*:1097–1105.
3. Freeman J, Simoncelli EP (2011) Metamers of the ventral stream. *Nat Neurosci* 14(9):1195–1201.
4. Willenbockel V, et al. (2010) Controlling low-level image properties: The SHINE toolbox. *Behav Res Methods* 42(3):671–684.
5. Konkle T, Caramazza A (2013) Tripartite organization of the ventral stream by animacy and object size. *J Neurosci* 33(25):10235–42.
6. Freeman J, Ziemba CM, Heeger DJ, Simoncelli EP, Movshon JA (2013) A functional and perceptual signature of the second visual area in primates. *Nat Neurosci* 16(7):974–81.
7. Okazawa G, Tajima S, Komatsu H (2015) Image statistics underlying natural texture selectivity of neurons in macaque V4. *Proc Natl Acad Sci U S A* 112(4):E351-60.