A 2D Gabor-wavelet baseline model out-performs a 3D surface model in scene-responsive cortex

Anna Shafer-Skelton[1,2], Timothy F. Brady[1], & John T. Serences[1,3,4]

[1] Department of Psychology, University of California, San Diego, La Jolla, USA
[2] Department of Psychology, University of Pennsylvania, Philadelphia, USA
[3] Neurosciences Graduate Program, University of California, San Diego, La Jolla, USA
[4] Kavli Institute for Brain and Mind, University of California, San Diego, La Jolla, USA

Correspondence:
Anna Shafer-Skelton
ashafers@upenn.edu

Short title: 2D Gabor-wavelet model out-performs 3D surface model in scene-responsive cortex)

Abstract

Understanding 3D representations of spatial information, particularly in naturalistic scenes, remains a significant challenge in vision science. This is largely because of conceptual difficulties in disentangling higher-level 3D information from co-occurring features and cues (e.g., the 3D shape of a scene image is necessarily defined by spatial frequency and orientation information). Recent work has employed newer models and analysis techniques that attempt to mitigate these in-principle difficulties. For example, one such study reported 3D-surface features were uniquely present in areas OPA, PPA, and MPA/RSC (areas typically referred to as 'scene-selective'), above and beyond a Gabor-wavelet baseline ("2D") model. Here, we tested whether these findings generalized to a new stimulus set that, on average, dissociated static Gabor-wavelet baseline ("2D") features from 3D scene-surface features. Surprisingly, we found evidence that a Gabor-wavelet baseline model better fit voxel responses in areas OPA, PPA and MPA/RSC compared to a model with 3D-surface information. This raises the question of whether previous findings of greater 3D information could have been due to a baseline condition that didn't model some potentially critical low-level features (e.g., motion). Our findings also emphasize that much of the information in "scene-selective" regions—potentially even information about 3D surfaces—may be in the form of spatial frequency and orientation information often considered 2D or low-level, and they highlight continued fundamental conceptual challenges in disentangling the contributions of low-level vs. high-level features in visual cortex.

Introduction

A large body of evidence in vision science argues that anatomically distinct regions of cortex are selective to and preferentially process images of natural scenes vs. those of discrete visual objects [1,2]. In particular, early work suggested that these "scene areas" were especially sensitive to configurations of major surfaces like walls and floors that define the 3D shape of a space [e.g., 2–4]. However, less is known about the format of this information or how such an anatomical distinction between scene and object processing may impact cognitive processes such as visual attention and memory. The nature of visual scene representations is challenging to study because of difficulty in disambiguating representations of scene surfaces per se both from representations of more global scene properties (like "openness" or "navigability") and from low-level features (like orientation or spatial-frequency information). Three well-studied regions of cortex offer fruitful areas to target in asking about the contributions of each of these possibilities (surfaces, global properties, low-level features) to scene representations: the occipital place area (OPA), parahippocampal place area (PPA), and retrosplenial complex/medial place area (RSC/MPA). All are characterized by higher responses to scene images than other stimuli, and they seem to contain information relevant to navigating through the 3D world [5–7].

Do these scene selective areas contain information about a scene's surfaces and 3D spatial layout per se rather than other visual properties of the scenes? One important challenge to addressing this is differentiating low- to mid-level visual features from the 3D information that they ultimately support. Scene-selective cortical areas could be responsive to configurations of 3D surfaces per se, but these tend to co-occur in the natural world with cues such as patterns of orientations and spatial frequencies. Foundational work on scene areas has employed hypothesis-driven stimulus conditions such as preserving the same surfaces in a scene while selectively

disrupting their configuration [e.g., 2]. These manipulations cannot (and don't pretend to) completely remove the influence of orientation and spatial frequency, since these basic features are the building blocks of all more complex images. Instead, they focus on finding a theoretically interesting level of tolerance to low-level feature variation: after finding higher PPA responses even for unfurnished rooms compared to isolated objects from those rooms, they sought to test whether the presence of a coherent configuration of surfaces was the critical aspect of the stimulus driving responses. They created two conditions: (1) fractured images (surfaces cut apart and separated by white space) vs. (2) fractured + re-arranged surfaces: the same fractured surfaces, but re-arranged within the image to prevent them being perceived as a 3D space. They found that the fractured surfaces, which still had a recognizable 3D structure, drove PPA responses similarly to the empty scenes – at least in terms of overall univariate magnitude averaged over the entire region. In contrast, univariate responses to the fractured + re-arranged surfaces were significantly lower than both real scenes and fractured scenes. While low-level features were not perfectly teased apart from the features of interest—e.g., the fractured rooms still contained edges oriented the same way in similar parts of the image—, these features have a close relationship to surface information in everyday life. Thus, this comparison was taken as a powerful demonstration that PPA cares about configurations of surfaces, insofar as intact 3D structure – even without the walls being connected – led to greater univariate activity in PPA than a scrambled version that broke 3D structure. Notably, however, this study – and similar early work [reviewed by 8] did not attempt to understand the representation of the stimuli in individual voxels or patterns of voxels, instead examining only which stimuli tended to lead to overall activation.

More recent studies have sought to compare spatial information of interest with the coarsely localized orientation and spatial frequency information captured by a GIST model [9], often thought of as lower-level visual information. For example, Bonner & Epstein [6] focused on the representational structure present in the pattern of OPA activity, showing that OPA in particular captures information about navigational affordances in scenes. Henriksson, Mur, & Kriegeskorte [10] find evidence for an OPA representation of 3D surfaces of a scene beyond GIST features. A notable aspect of both of the above findings was the surprising amount of information corresponding to GIST features (even more than the affordance features in Bonner & Epstein), perhaps reflecting the close relationship observed between GIST features and spatial and/or category information that may be important for scene-selective areas [9,11,12]. Given this close relationship, considering GIST features to be a "low-level" baseline may result in some meaningful information about the 3D shape of a scene being discounted—a possibility that complicates interpretations specifically when GIST features perform well.

The results from another study [13] avoided this conceptual difficulty by showing that a novel encoding model of 3D scene surfaces better predicted individual voxel responses compared to GIST and other baseline models. Their 3D model captured spatial characteristics of major surfaces (distance and orientation in 3D space) in videos of moving scenes (see Figure 1 for stimulus example). The performance of this model was compared against the best-performing of three baseline models, each capturing information about orientation and spatial frequency in static images based on the videos. The high performance of the 3D model compared even to the best of the three baseline models was a strong endorsement of this new model—while the GIST model had performed close to or better than the models of interest in the studies above [6,10], Lescroart & Gallant's 3D model performed better even than the gabor wavelet baseline model,

which itself performed more than twice as well as the GIST model. Gabor wavelet models have previously been successful at predicting neurophysiological and fMRI data from EVC, and they have also served as baseline models for higher-level visual cortical representations of static image stimuli [e.g., 14]. However, for moving stimuli, this static Gabor baseline model raises concerns: while the video stimuli contain motion parallax as a depth cue, the baseline model does not model low-level motion cues correlated with this parallax (e.g., horizontal motion in the lower plane of the screen that could signal surfaces closer in space), as it ignores motion information. This leaves open the possibility that the 3D info that was observed in OPA/PPA/RSC is not 3D info *per se*, but instead may reflect variance that is attributable to un-modeled low-level motion information that co-varies with 3D scene surface features in moving video stimuli. Addressing this directly in that data set would require comparison to a baseline model that accounts for motion energy as well as 2D orientation structure.

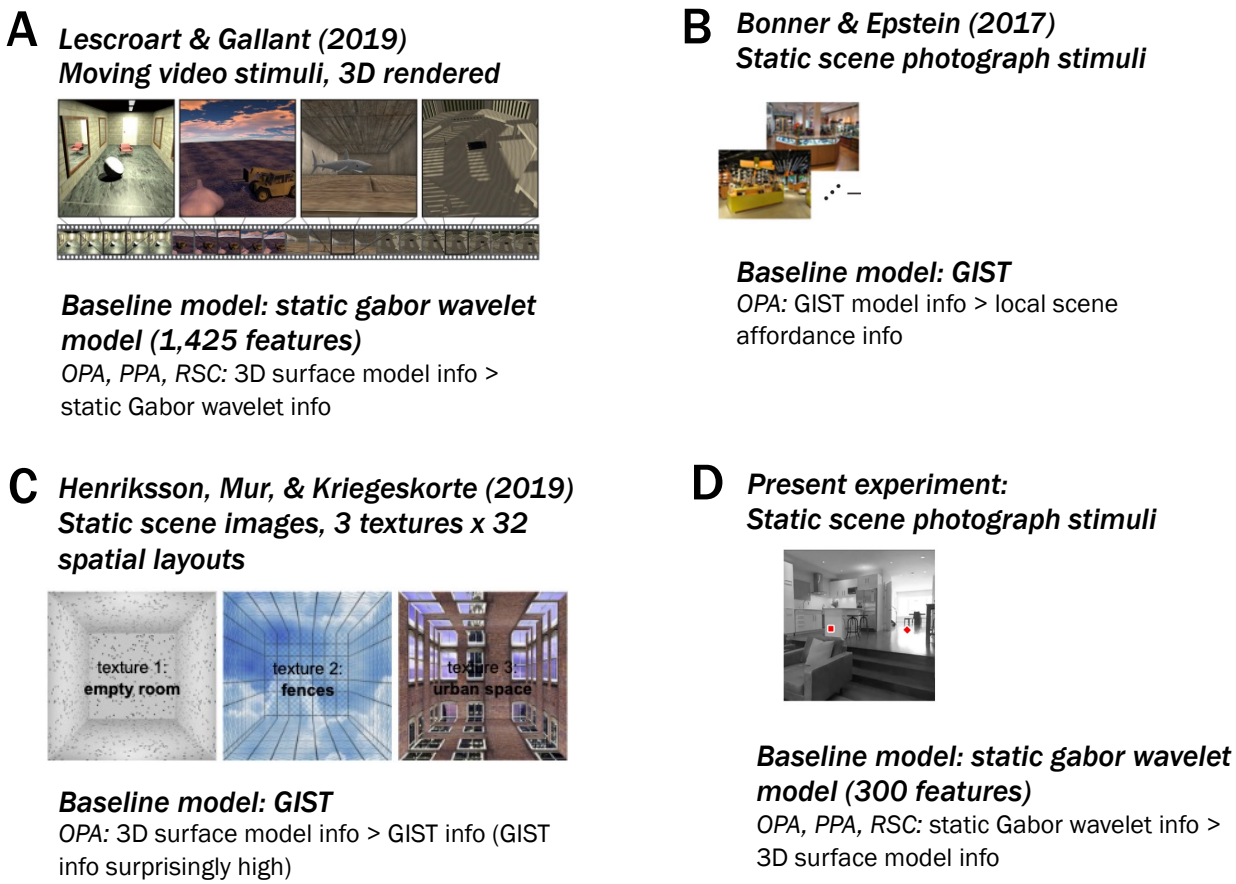


Figure 1: Stimuli from 3 related studies (A-C) as well as the present experiment. (A) Lescroart & Gallant (2019) used 3D-rendered moving video stimuli, used a static gabor wavelet baseline model, and in OPA, PPA, and RSC found 3D information beyond their baseline model. (B) Bonner & Epstein (2017) used static scene photograph stimuli, a GIST (static) baseline model, and in OPA found greater GIST than local scene affordance information as well as some information uniquely attributable to local scene affordances. (C) Henriksson, Mur, & Kriegeskorte (2019) used 3D-rendered static scene image stimuli, a GIST baseline model, and in OPA found greater 3D surface model info, along with surprisingly high GIST information. (D) The present experiment uses static scene photograph stimuli, a static gabor baseline model, and in OPA, PPA, and RSC found that a Gabor-wavelet baseline model accounts for more variance than a 3D surface-model.

In the present study, rather than attempt to expand a baseline model to account for the complexities of video stimuli, we instead removed any contribution of low-level motion information to voxel responses by using static photograph stimuli. While this sacrifices one aspect of "naturalness" of the stimulus set, removing motion energy allows us to use a static

Gabor model – similar to Lescroart and Gallant's – as a baseline model, with this model now encompassing the set of low-level features we wish to discount (orientation and spatial frequency).

We then compared this baseline model to a 3D surface model in OPA/PPA/RSC. If the use of static images and a Gabor baseline model leaves 3D information in OPA/PPA/RSC relatively unaffected, then our results would suggest that prior analysis approaches are robust to the in-principle difficulties described above. However, if the Gabor wavelet model now performs equally to or better than the 3D model, the approach of taking a gabor wavelet model as a proxy for 2D information to be discounted may not directly answer the question of whether there is 3D information *per se* in areas like OPA/PPA/RSC.

To model 3D surface information, we computed two sets of features capturing the distance to and 3D directions of surfaces in each image: first, we used a model capturing summaries of these 3D characteristics across the entire image (as in Lescroart & Gallant, 13; "3D-global model"); second, we used a version summarizing these 3D features within each quadrant of the image, preserving more fine-grained spatial information ("3D-quadrant-based model"). While we also manipulated the task-relevance of the 3D spatial structure of the scene and the spatial extent presented small probe shapes used to keep the subjects engaged (see *Methods* for details), we saw no indication of any effects of these manipulations on our results and don't focus on them in the present paper.

Thus, we investigate the generalizability of previous 3D scene information in OPA/PPA/RSC in two key ways. First, our stimuli consisted of real-world photographs (instead of the dynamic 3D-rendered scenes used in Lescroart & Gallant, 2019), allowing the baseline Gabor wavelet pyramid model to encompass the relevant 2D features to be discounted, since

motion was no longer present in the stimuli. In addition, we selected the set of static image stimuli using an algorithm to ensure that across all images, the feature sets of interest are highly separable.

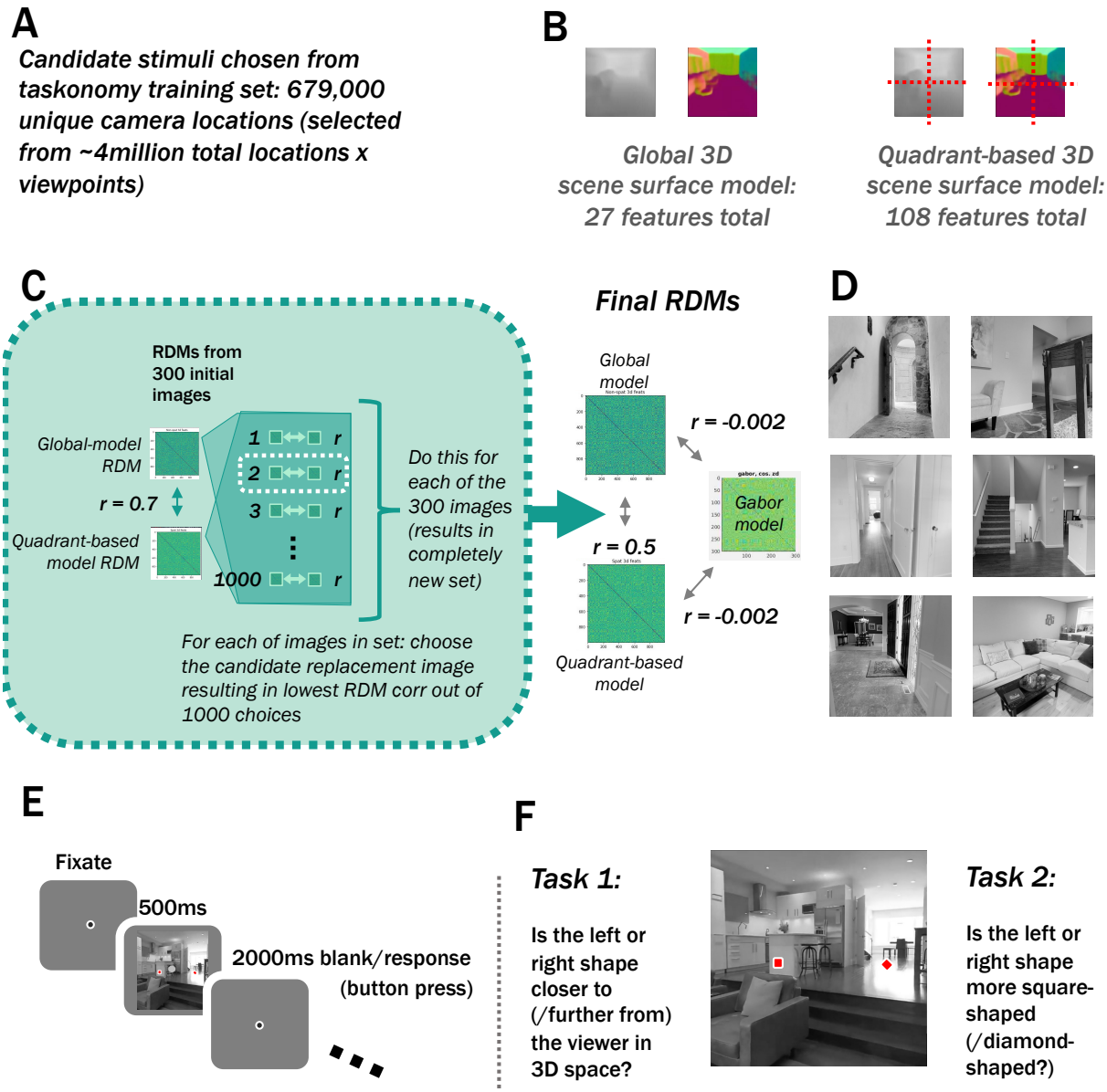


Figure 2. Methods. (A) Candidate stimuli were drawn from the taskonomy training set. (B) Feature sets were computed similarly to Lescroart & Gallant (2019), from Euclidean-distance and surface-normal annotations included in the Taskonomy [15] training set. (C) The 300 stimulus images were chosen to minimize 3D-global- vs. 3D-quadrant-based-model RDMs. An initial set of stimulus images (chosen similarly to Groen et al., [16]) had a correlation of 0.7. Next, we replaced each of the 300 stimulus images with the image from 1000 random choices that resulted in the lowest RDM correlation. (D) Example stimulus images. (E) Example trial. (F) Tasks. On alternating runs, participants either answered 1) whether the left or right shape was

closer to (farther from) the viewer in 3D space or 2) whether the left or right shape was more square-shaped (diamond-shaped).

Results

Figure 3 shows the fit of each model in each visual region.

EVC performance serves as a baseline, replicating previous results: As expected based on prior work in other labs, we observed that the gabor model [17] and the 3D-surface model predicted voxel-wise activity above chance [10,13] in early visual areas (all *p*s < 0.0003, uncorrected; Table S2A). We also found similar magnitudes of cross-validated prediction performance (see Figure 3 in [13]).

Stronger Gabor-wavelet-baseline vs. 3D representations in scene areas; no unique evidence for 3D representations: While we found that prediction performance was significantly above chance for all three models in areas OPA/PPA/RSC (all *p*s < 0.014; see Figure 3A and Table S2A), we surprisingly observed that the gabor-wavelet-baseline model outperformed the 3D models (Figure 3A; all ps < 0.0026). This result prompted us to apply a 3D-vs.-Gabor-wavelet-baseline variance-partitioning analysis in these areas to determine the unique contributions of each set of features [13]. We adopted this approach because the small amount of 3D information we found could either be due to 3D features per se, or it could be due to imperfect separability of 3D vs. Gabor-wavelet-baseline features in our stimulus set, which would mean that information that is genuinely Gabor-wavelet related could contribute to a small degree to the fit of the 3D model. Surprisingly, we found that only the *Gabor-wavelet baseline* features uniquely predicted responses in these regions (unique Gabor-wavelet-baseline info in variance partitioning vs. 3D-quadrant-based features: OPA: p = 0.0006; PPA: p = 0.0002; RSC: p = 0.026; unique Gabor-wavelet-baseline info in variance partitioning vs. 3D-global features:

OPA: p = 0.0018; PPA: p = 0.0005; RSC: p = 0.053). In this analysis, shared variance represents information not uniquely attributable to either model. In each area, there was significant *shared* 3D-quadrant-based/ Gabor-wavelet-baseline (OPA: p = 0.0019; PPA: p = 0.0004; RSC: p = 0.026) and *shared* 3D-global/Gabor-wavelet-baseline (OPA: 0.0069; PPA: 0.0018; RSC: 0.0048) representations, with no significant evidence for uniquely 3D representations (3D-global model: all *p*s > 0.13; 3D-quadrant-based model: all *p*s > 0.41; Figure 3B and C and Table S2 B and C). Note that RSC results follow the same pattern as the OPA/PPA results (significant unique-Gabor and shared variance, without significant unique-3D variance) but are not as robust, consistent with previous work. In OPA and PPA, the unique variance explained by the Gabor-wavelet-baseline model was significantly greater than either the 3D-quadrant-based model (ps < 0.0008) or the 3D-global model (ps < 0.004). In addition, unique Gabor-wavelet-baseline model variance was greater than the shared variance not uniquely attributable to either model: unique Gabor > shared variance between Gabor-wavelet-baseline and the 3D-quadrant-based model (ps < 0.0027) and unique Gabor-wavelet baseline > shared variance between the 2D gabor-wavelet model and the 3D-global model (ps < 0.0061). Again, RSC showed a similar pattern of greater Gabor-wavelet-baseline performance but was less reliable overall (Figure 3B and C and Table S2 B and C).

Next, we investigated whether these results were robust to methodological decisions such as voxel selection. First, plotting each model's performance as a function of selecting different numbers of voxels in each ROI, we find a strikingly consistent pattern of relative model performance, arguing against this pattern being an artifact of voxel selection procedure (see *Methods: Regions of Interest*). Note that while, in hindsight, our pre-determined voxel counts may have lowered the *overall magnitude* of performance for all models, a more conservative

cutoff would not qualitatively change our results. Next, a sliding window analysis (Figure S3) shows that these results are also consistent across timepoints. Finally, while we collapse across two tasks in our main analyses (one for which the 3D structure is relevant and one for which it is not), the 3D-relevant task data alone show an almost identical pattern of model performances in all ROIs (Figure S2).

Together, these results show a pattern of higher Gabor-wavelet-baseline performance compared to 3D scene-surface models in OPA/PPA/RSC. This pattern of results was robust to changes in several methodological decisions, participants' task, and the variation of the 3D model that we chose.

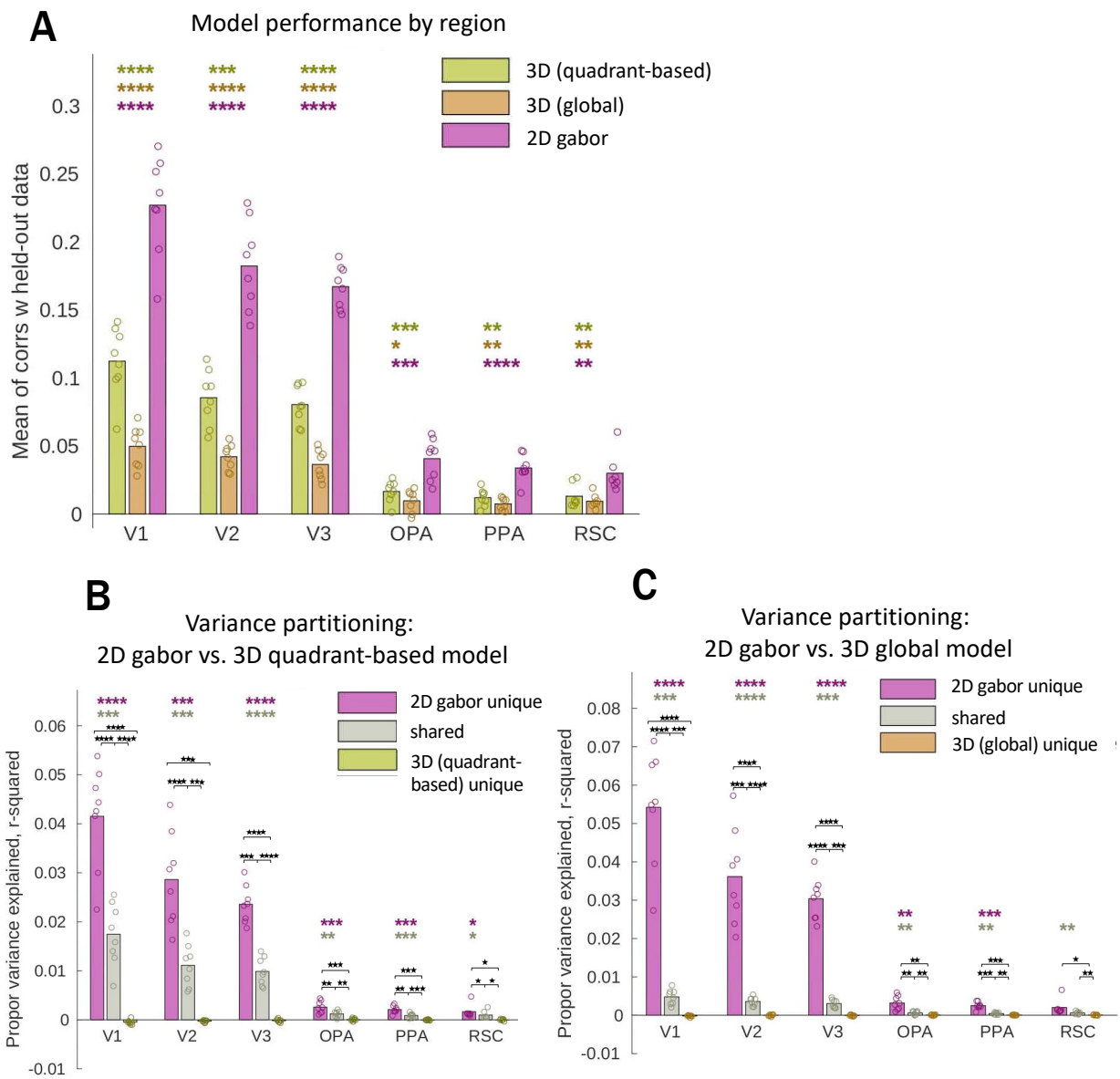


Figure 3. Dots represent individual participant data. Stars denote significant two-tailed effects in the positive direction (**** = p<0.0001; *** = p<0.001; ** = p<0.01; * = p<0.05, uncorrected). Colored stars correspond to individual tests, and brackets indicate paired tests. Note that magnitudes across ROIs are not interpretable, but patterns within ROIs are. (A) Results collapsed across all task data. (B) Variance partitioning results for "2D" Gabor-wavelet-baseline model vs. 3D-quadrant-based model. (C) Variance partitioning results for "2D" Gabor-wavelet-baseline model vs. 3D-global model.

Discussion

The present work investigated whether 3D scene-surface information in OPA, PPA, and RSC (e.g., Lescroart & Gallant, 2019) is robust to comparisons with a baseline model that includes the relevant low-level information (in this case, static orientation and spatial frequency info) present in the stimuli (in this case static scene photographs). We included two variations of a 3D scene-surface model: *3D-global* (summarizing across the entire image) and *3D-quadrant-based* (summarizing across quadrants of the image), comparing each against a baseline *Gabor-wavelet model* capturing orientation and spatial-frequency information. We found higher prediction performance for the Gabor-wavelet-baseline model compared to each of the 3D-scene-surface models, in each scene-responsive cortical area. In each of these areas, we also found evidence for unique Gabor-wavelet-baseline representations (representations uniquely attributable to this model in a variance partitioning analysis) but not unique 3D representations. This appears contrary to the dominant claim about OPA in particular [5–7,13].

No significant evidence for representations uniquely attributable to 3D scene surface information: Our variance partitioning analysis revealed no evidence that 3D scene surface features uniquely explained activity in scene-selective cortex. These results stand in contrast with the conclusions reached by Lescroart & Gallant [13], who examined responses to moving videos of scenes. As suggested in the introduction, the higher performance of the 3D-surface-model in their work could be partially attributable to contributions from low-level motion cues that could not be accommodated by the static Gabor-wavelet model they use as a baseline. This is because, just like the static Gabor information Lescroart & Gallant [13] discounted, motion information can also correlate with 3D surface information present in stimulus videos. While we sacrificed

the ability to model more naturalistic moving stimuli, our use of static photographs removed motion information from the stimulus set, thus allowing us to use a static Gabor-wavelet baseline model that accounted for the low-level featural information in the images (orientation and spatial frequency). With this change, we observed that the Gabor-wavelet-baseline model generally outperformed the 3D model, even in OPA/PPA/RSC.

In addition to using moving stimuli, the stimuli used by Lescroart & Gallant [13] also employed lighting and texture differences that may disrupt the *across-stimuli* relationship between low-level visual features and 3D surfaces. The importance of this issue is illustrated by imagining that, across images, rightward-facing surfaces usually co-occur with down-right oriented lines in one part of the image, and/or up-right orientations in a different part of the image (etc. for other surface orientations). The more of these types of co-occurrence, the more difficult it would be for regression models to disentangle the variance explained by the gabor and the 3D-surface models. By rendering their scenes using a 3D modeling software, Lescroart & Gallant [13] sought to introduce high-contrast shadows and other visual features to disrupt this co-occurrence between 3D surfaces and low-level visual features (i.e., oriented edges). While we took a different approach—selecting a set of naturalistic scene photographs from a large image database—, two aspects of our results argue that a failure to reduce co-variation between feature sets via stimulus selection was not a reason for the differing results, perhaps because our stimulus selection procedure capitalized on naturally occurring disruptions in these co-variations. First, we find relatively low shared variance between these models (variance not uniquely attributable to either model), even lower than Lescroart & Gallant [13]. Second, even if we imagined all shared variance being attributed uniquely to the 3D surface model (i.e., imagine stacking the gray bars in Figure 3 on top of the chartreuse (3b) or orange (3c) bars), the Gabor-

wavelet baseline model would still explain more unique variance and perform better than the 3D surface model (the pink bars would still be higher in OPA/PPA/RSC than the gray + chartreuse and the gray + orange).

We also investigated the robustness of our results to several different methodological decisions. First, when varying ROI size, we still see greater gabor-wavelet performance than 3D-model performance, even within individual participants (Figure S4). While our main results investigate voxel responses within a narrower time window than Lescroart & Gallant's [13] and without the added flexibility of FIR features, our pattern of results is consistent across a wide range of timepoints (Figure S3). Finally, we see the same pattern when only using data for which participants performed a task requiring attention to the 3D scene structure (Figure S2) as we do using data collapsed across tasks (main results).

It is also worth noting that both the present study and Lescroart & Gallant [13] fit models using cross-validation, making the results robust to overfitting. Further, while one of our 3D models was smaller than Lescroart & Gallant's due to the selection of fewer distance bins (9 vs. 3; resulting in 27 and 108 features in our models vs. 91 features in Lescroart & Gallant's), it seems unlikely that this change could be a major factor in our dramatic reversal of results, since Lescroart & Gallant's data [13; Figure 3A] showed that their surface-orientation model, without any distance information, performed nearly as well as their full 3D model. In addition, we also reduced the number of features in our Gabor-wavelet-baseline model to 300 (vs. 1,425 in Lescroart & Gallant's). As our stimuli consisted primarily of indoor scenes, we cannot completely rule out a critical role of the sky channel in driving the success of Lescroart & Gallant's 3D models. However, in their Figure 6, the first two PC's of voxel responses do not appear to be trivially capturing the presence of sky pixels [13]. Thus we don't believe a lack of

generalization to indoor scenes to be a likely reason for the higher performance of the Gabor-wavelet-baseline than the 3D-global and the 3D-quadrant-based models in our data set.

While our conclusions may at first seem at odds with the field's present understanding of scene-responsive cortex, two studies in particular have explicitly used a baseline model capturing spatial frequency and orientation information (a gist model [9]) and shown results compatible with ours. First, while investigating local-scene-affordance information (where it would be possible to traverse the 3D space depicted in a scene), Bonner & Epstein [6] found a similar pattern of higher performance for GIST features than their local-affordance (arguably 3D) information. While they did still find a small amount of unique local-affordance information, the presence of this information is compatible with our results in light of Lescroart et al.'s [13] data showing a Gabor wavelet model performing over twice as well in OPA/PPA/RSC as a GIST baseline model. This model difference also potentially affects Henriksson et al. [10], which also used a GIST model as their baseline model—such a higher-performing Gabor-wavelet baseline model could be expected to out-perform their 3D information as well.

Overall, we believe that when performing model comparisons between richer, higher-level models (like 3D surfaces and affordances) and simpler low-level models (like GIST, or Gabor wavelet), it is critical that baseline models include all relevant types of low-level information (e.g., motion energy when using videos). This is easier to do when using simple, non-moving stimuli, such those employed in the present study. However, future work could leverage the obvious advantages of more naturalistic stimuli such as those employed by Lescroart and Gallant [13], achieving even greater insight by building successively more complex models to account for different aspects of 3D processing.

Implications for the visual system: The present pattern of results, considered in conjunction with other data, suggests that when Gabor-wavelet orientation and spatial frequency features (often considered 2D, or low-level) are used as a baseline, there is less detectable unique 3D surface information in scene-selective cortex. However, this doesn't mean that scene areas can't still be *for* scenes—importantly, there is evidence that patterns of orientation and spatial frequency may actually serve a critical role in processing scene information: first, the distinction between "2D" and "3D" is not as binary as it may seem from these labels— 2D orientations/spatial frequencies are computationally sufficient to infer global scene properties related to 3D layout [18], and human behavior does indeed seem to rely on these features in scene processing tasks [11]. Thus, in a baseline-model selection procedure in which the best-performing "2D" model becomes the final baseline [13], there may actually be some meaningfully 3D information discounted because it was captured by the baseline model. This realization highlights a critical and complementary role for baseline conditions motivated by theoretical considerations, and further experimentation that can provide more insight into the most theoretically interesting amount/type of information to discount.

In this work, we originally sought to use a task manipulation to gain insight into a special characteristic of scene representations: the reduced attention with which some types of scene information can apparently be processed [19,20]. The results discussed here highlight the lack of agreed-upon markers of scene-specificity: a pre-requisite to investigating this topic and an area that will be critical to address in understanding our behavior within natural environments.

Conclusion: The present work further examines the presence of 2D and 3D representations in areas proposed to part of scene-selective cortex (PPA/OPA/RSC) while participants viewed naturalistic scene-photograph stimuli. In contrast to some previous work, we

find that, when using a baseline model that encompasses the cues to 3D surface representations, there is a marked reversal such that Gabor-wavelet baseline ("2D") models compare favorably to models of 3D information in scene-responsive cortex, even when accounting for potentially shared variance. This work highlights the difficulty in drawing conclusions about separability of 3D (or higher-level) representations from the patterns of orientation and spatial frequency that underlie our representations of them. This motivates future, more targeted investigations of exactly which information is "3D" or meaningfully related to higher-level properties of scenes.

Methods

Participants: 8 participants (6 female, 2 male) between the ages of 25 and 31 participated in the experiment. The protocol was approved by the Institutional Review Board of UCSD, and all participants gave informed written consent. Each participant completed a behavioral training session lasting approximately 1.5 hours, followed by a retinotopic mapping session and three fMRI sessions where they performed the main task. Each scanning session lasted approximately 2 hours. Participants were compensated $15/hr for the behavioral session and $25/hr for each of the three scanner sessions.

fMRI data acquisition and pre-processing: fMRI data collection was completed at the UC San Diego Keck Center for Functional Magnetic Resonance Imaging, on a General Electric (GE) Discovery MR750 3T scanner. Functional echoplanar imaging (EPI) data were collected using a Nova 32-channel head coil (NMSC075-32-3GE-MR750) and the Stanford Simultaneous Multi-Slice (SMS) EPI sequence (MUX EPI), with a multiband factor of 8. This resulted in 9 axial slices per band (72 slices total; 2mm^3 isotropic voxels; 0mm gap; matrix = 104x104; FOV = 20.8cm; TR=800ms; TE = 35ms; flip angle = 52°; in-plane acceleration = 1).

We acquired a high-res T1 anatomical scan in the same session (GE ASSET on a FSPGR T1-weighted sequence; 1x1x1 mm^3 voxel size; 8136ms TR; 3172ms TE; 8° flip angle; 172 slices; 1mm slice gap; 256x192cm matrix size), corrected for inhomogeneities in signal intensity using GE's Phased array uniformity enhancement (PURE) method. This was used for segmentation, flattening, and delineation of retinotopic visual areas.

Preprocessing was completed using Freesurfer and FSL (available at https://surfer.nmr.mgh.harvard.edu and http://www.fmrib.ox.ac.uk/fsl). We used Freesurfer's recon-all utility [21] to perform cortical surface gray-/white-matter segmentation of each subject's anatomical scan. These segmentations were then used to define cortical surfaces on which we delineated retinotopic ROIs used for subsequent analyses in EVC (see *Regions of Interest*). We used Freesurfer's manual and automatic boundary-based registration tools [22] to generate transformation matrices that were then used by FSL FLIRT [23,24] to co-register the first volume of each functional run into the same space as the anatomical image. Motion correction was performed using FSL MCFLIRT [24], without spatial smoothing, with a final sinc interpolation stage, and with 12 degrees of freedom. Finally, slow drifts in the data were removed using a high-pass filter (1/40 Hz cutoff). No additional spatial smoothing was performed for main task runs. After pre-processing, we z-scored each voxel's timeseries within each run and epoched the data based on the start time of each trial. Because of the short ITI and fast event-related design, we averaged responses between 2.4 and 4.8 seconds after stimulus onset to use as each trial's data for our main analyses, unless specified otherwise.

Procedure and experimental conditions: For all task runs described here, stimuli were projected onto a semi-circular screen 21.3cm wide and 16cm high, fixed to the inside of the scanner bore just above the participant's chest. The screen was viewed through a mirror attached to the head coil, from a viewing distance of 49 cm. After taking into account the shape of the screen and the square stimuli, this resulted in a vertical extent of approximately 18.1° (max vertical eccentricity of 18.1°/2). The background was a mid-gray color, with a darker gray placeholder outline marking the location of the square stimuli between stimulus presentations.

The fixation point was a black rounded square of 0.2° with a white outline and was on the screen throughout each run.

We collected a total of 12 experimental runs (514 TRs each; 6 minutes 52 seconds long) in each of the three sessions, for a total of 36 runs per participant. Runs alternated between two tasks (see *Tasks* section below), with task order counterbalanced across participants. Within each run, three mini-blocks corresponded to narrow, medium, or broad regions of the scene ("spatial spread") that target shapes could appear in (see *Dot Selection* below). The order of mini-blocks within each run was counterbalanced within participants to minimize order effects of spatial-spread conditions. Four combinations of response mappings were counterbalanced across participants along with task order, for a total of one combination per participant.

We used a fast event-related design, with each stimulus presented for 500ms on each trial, with a 2000-ms un-jittered response window/ITI.

Stimuli and Tasks: *Main task:* On each trial, a pair of red shapes (one square and one diamond) was superimposed on a grayscale scene photograph. In *distance judgment* runs, participants were instructed to respond whether the left or the right shape was on the part of the scene that would be closer to (/farther from) the viewer in three-dimensional depth if viewed in real life [25,26]. In *shape judgment* runs, the participant was instructed to respond whether the left or right shape was a square (/diamond). Participants responded "left" (square/diamond/closer/farther) using their index finger and "right" using their middle finger. All responses were made with their right (dominant) hand. Response mappings were consistent throughout the entire session for each participant, including during the 90-minute behavioral practice session. To ensure that any task differences were not due to different difficulties of

tasks, the diamond and square could be parametrically adjusted to be more rounded, allowing us to continuously adjust the difficulty of the task to match the distance judgment task. Dot pairs were selected to keep participants off ceiling in the distance judgment runs so that difficulty could be accurately matched. (See Figure S1 for participant performances across task.) By choosing two tasks that depended on judgments of two dots, we aimed to avoid task effects that could be trivially explained by small eye movements or covert spatial attention.

Functional localizer task: During the localizer task, participants viewed blocked presentations of images of scenes, objects, faces, bodies, scrambled images (images divided into small square pieces and randomly re-arranged), as well as a fixation baseline condition. They pressed a button when they detected trials in which the same image was identical to the image immediately previous (one-back task). Stimulus presentation code was adapted from code from the Grill-Spector Lab.

Stimulus selection: We include 300 stimulus images from the Taskonomy training set selected to maximally differentiate between the 3D-global and 3D-quadrant-based scene-surface models. The Taskonomy training set is made up of 4.5 million scene images from hundreds of unique buildings, with a range of different camera locations in each building. We first subsetted this data set by selecting one image from almost all unique camera locations, resulting in 679,000 images in our starting set. We began with an iterative approach similar to Groen et al.'s [16], creating representational similarity matrices (RDMs) based on the cosine distance between features of each image. Iterating through randomly chosen stimulus *sets* resulted in a plateaued correlation of ~0.7 between 3D-global and 3D-quadrant-based feature RDMs. We next iterated through each of the 300 images in this initial set, testing RDM correlations for 1000 potential

replacement images and using the replacement that resulted in the lowest RDM correlation of the set. After replacing each of the 300 images once, we were left with a correlation of 0.5 between 3D-global and 3D-quadrant-based feature RDMs. Although this procedure did not explicitly orthogonalize either 3D feature set against gabor features, we ended up with smaller RDM correlations (3D-global vs. gabor: -0.002) than Lescroart et al.'s [13] validation data set (0.274). This pattern was robust to choice of distance metric (see Table S1).

Dot selection: For each stimulus image, we chose locations of 3 pairs of superimposed dots (the square- and diamond-shaped dots described in the *Tasks* section above): one each within a narrow, medium, and broad spread of potential dot locations. We did this using a semi-automated Matlab script, indicating the allowed region of the image for each spatial spread condition and prompting the selector to (if possible) select via mouse click one potential set in which the correct answer was "left closer" and one in which the correct answer was "right closer". Feedback was given in the command window after each selection to help the selector choose dot locations with roughly similar average differences in distance across dot pairs in the narrow, medium, and broad spatial spread conditions. To choose the final set from these manually selected options, a Matlab script chose from these options to contain an equal number of left-correct and right-correct dot pairs and verified that the narrow, medium, and broad conditions didn't differ substantially in the depth difference between pairs of dots.

Feature sets: Gabor wavelet baseline model: The *Gabor-wavelet model* contained 300 features, each a combination of one of 4 spatial frequencies (0, 2, 4, 8, and 16 cycles per image) and one of 4 orientations (0, 45, 90, or 135 degrees). The stimulus image was tiled by a grid of

gabor wavelets for each spatial frequency. We also compared fitting success in one subject using a 1,425-feature gabor model, achieving similar results as with the 300-feature model. Thus, we opted to use the smaller model for all main analyses. These features were generated using code from https://github.com/gallantlab/motion_energy_matlab.

Feature sets: 3D scene surface models: 3D-g*lobal* and 3D-*quadrant-based scene surface features* were computed using the corresponding ground-truth distance and surface-normal (surface direction) image maps included with our stimulus images, which were selected from the training set for the Taskonomy [15] neural networks. To generate features from these maps, we adapted code from Lescroart & Gallant [13], first grouping pixels in the distance maps into 3 bins, spaced so that distance bins were roughly equally represented across images. Using the surface-normal maps, we grouped each pixel into one of 9 surface-direction bins: forward, upward, downward, leftward, rightward, up-left, up-right, down-left, and down-right. In the *3D-global scene-surface model*, we counted the proportion of pixels in each image falling into each combination of the 3 distance and 9 surface-direction bins for a total of 27 features. In the *3D-quadrant-based scene-surface model,* we next divided the image into 4 quadrants, with each of the 108 resulting features corresponding to a combination of image quadrant x surface direction x distance bin. Before model fitting, features were z-scored.

Ridge regression: We fit models using ridge regression, with nine cross-validation folds. For the analyses collapsing across task, each of the 9 left-out sets was made up of 4 of the 36 runs (two from each task). For the scene-task-only results, each of the left-out sets was made up of 2 of the 18 same-task runs. To choose ridge parameters, we nested 10 inner cross-validation folds within the training data of each main fold and chose the lambda predicting the highest r-

squared among the inner cross-validation folds. This lambda was then used to fit all of the data in the training portion of the main fold. There were 13 possible ridge parameters: 0, as well as 12 values logarithmically spaced between 10^{-2} and 10^{5}. Early analyses on pilot data showed that this ridge parameter selection matched performance using a more extensive set of ridge parameter choices.

Our main analyses report prediction performance, first averaged across the 9 main cross-validation folds, then averaged across all voxels in an ROI. Statistical significance was computed via permutation tests, shuffling the test data 10,000 times relative to the predicted data before completing the same process of averaging prediction performance across cross-validation folds and across voxels in each ROI. This generated permuted values corresponding to each ROI and participant, which we then used to calculate a null distribution of t-statistics across participants. We tested individual model performances against zero by calculating a t-statistic from the actual data and comparing that to the null distribution. We report two-tailed uncorrected p-values without assuming a symmetrical null distribution.

Regions of interest: For regions V1, V2, and V3, we followed previously established retinotopic mapping protocols [27–29]. Initial masks for areas V1, V2, and V3 were manually drawn based on retinotopic mapping data collected in a separate session, and candidate voxels were selected for further analyses based on a scrambled > baseline contrast (see Stimuli & Tasks: Functional Localizer Task, above), using a false discovery cutoff of $q < 0.05$. We used a combination of localizer runs from past experiments in our lab and the present experiment (stimuli from Grill-Spector lab), for a total of 4-10 functional localizer runs per subject. For OPA, PPA, and MPA/RSC, initial masks were manually drawn around contiguous clusters of

voxels in each subject's native space, including voxels with $q < 0.05$ for a scenes > objects contrast. VOI definitions were mutually exclusive. To ensure that any differences in performance across VOIs were not due to averaging over dramatically different numbers of voxels, we capped our main analyses to the 200 voxels with the strongest localizer signal in each bilateral VOI; we included VOIs with >= 75 voxels bilaterally, resulting in 7/8 participants with all VOIs defined and one with all defined except MPA/RSC (see Figure S4 for main analysis results as a function of number of voxels included). This participant is omitted in MPA/RSC analyses.

References

[1] Dilks DD, Julian JB, Paunov AM, Kanwisher N. The occipital place area is causally and selectively involved in scene perception. J Neurosci 2013;33:1331–6a. https://doi.org/10.1523/JNEUROSCI.4081-12.2013.

[2] Epstein R, Kanwisher N. A cortical representation of the local visual environment. Nature 1998;392:598–601. https://doi.org/10.1038/33402.

[3] Landau B, Lakusta L. Spatial representation across species: geometry, language, and maps. Current Opinion in Neurobiology 2009;19:12–9. https://doi.org/10.1016/j.conb.2009.02.001.

[4] Hermer L, Spelke ES. A geometric process for spatial reorientation in young children. Nature 1994;370:57–9. https://doi.org/10.1038/370057a0.

[5] Park J, Park S. Coding of navigational distance and functional constraint of boundaries in the human scene-selective cortex. Journal of Neuroscience 2020;40:3621–30. https://doi.org/10.1523/JNEUROSCI.1991-19.2020.

[6] Bonner MF, Epstein RA. Coding of navigational affordances in the human visual system. Proc Natl Acad Sci U S A 2017;114:4793–8. https://doi.org/10.1073/pnas.1618228114.

[7] Henriksson L, Mur M, Kriegeskorte N. Rapid invariant encoding of scene layout in human OPA. BioRxiv 2019:577064. https://doi.org/10.1101/577064.

[8] Epstein R. The cortical basis of visual scene processing. Vis Cogn 2005.

[9] Oliva A, Torralba A. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. Int J Comput Vis 2001;42:145–75. https://doi.org/10.1023/A:1011139631724.

[10] Henriksson L, Mur M, Kriegeskorte N. Rapid Invariant Encoding of Scene Layout in Human OPA. Neuron 2019;103:161-171.e3. https://doi.org/10.1016/j.neuron.2019.04.014.

[11] Brady TF, Shafer-Skelton A, Alvarez GA. Global ensemble texture representations are critical to rapid scene perception. Journal of Experimental Psychology: Human Perception and Performance 2017;5:0–17. https://doi.org/10.1037/xhp0000399.

[12] Oliva A, Torralba A. Building the gist of a scene: the role of global image features in recognition. Prog Brain Res 2006;155:23–36. https://doi.org/10.1016/S0079-6123(06)55002-2.

[13] Lescroart MD, Gallant JL. Human Scene-Selective Areas Represent 3D Configurations of Surfaces. Neuron 2019;101:178-192.e7. https://doi.org/10.1016/j.neuron.2018.11.004.

[14] Ratan Murty NA, Bashivan P, Abate A, DiCarlo JJ, Kanwisher N. Computational models of category-selective brain regions enable high-throughput tests of selectivity. Nat Commun 2021;12:5540. https://doi.org/10.1038/s41467-021-25409-6.

[15] Zamir A, Sax A, Shen W, Guibas L, Malik J, Savarese S. Taskonomy: Disentangling task transfer learning. IJCAI International Joint Conference on Artificial Intelligence, vol. 2019-Augus, 2019, p. 6241–5. https://doi.org/10.24963/ijcai.2019/871.

[16] Groen II, Greene MR, Baldassano C, Fei-Fei L, Beck DM, Baker CI. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. Distinct Contributions of Functional and Deep Neural Network Features to Representational Similarity of Scenes in Human Brain and Behavior 2017:207530. https://doi.org/10.1101/207530.

[17] Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature 2008;452:352–5. https://doi.org/10.1038/nature06713.

[18] Ross MG, Oliva A. Estimating perception of scene layout properties from global image features. Journal of Vision 2010;10:2.1-25. https://doi.org/10.1167/10.1.2.

[19] Alvarez G a., Oliva A. Spatial ensemble statistics are efficient codes that can be represented with reduced attention. Proceedings of the National Academy of Sciences of the United States of America 2009;106:7345–50. https://doi.org/10.1073/pnas.0808981106.

[20] Iris X, Groen IA, Ghebreab S, Lamme VAF, Scholte HS. The time course of natural scene perception with reduced attention. Journal of Neurophysiology 2016;115:931–46. https://doi.org/10.1152/jn.00896.2015.-Attention.

[21] Dale AM, Fischl B, Sereno MI. Cortical Surface-Based Analysis. NeuroImage 1999;194:179–94.

[22] Greve DN, Fischl B. Accurate and robust brain image alignment using boundary-based registration. NeuroImage 2009;48:63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060.

[23] Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. Medical Image Analysis 2001;5:143–56. https://doi.org/10.1016/S1361-8415(01)00036-6.

[24] Jenkinson M, Bannister P, Brady M, Smith S. Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. NeuroImage 2002;17:825–41. https://doi.org/10.1006/nimg.2002.1132.

[25] Sanocki T. Representation and perception of scenic layout. Cognitive Psychology 2003;47:43–86. https://doi.org/10.1016/S0010-0285(03)00002-1.

[26] Shafer-Skelton A, Brady TF. Scene layout priming relies primarily on low-level features rather than scene layout. J Vis 2019;19:14. https://doi.org/10.1167/19.1.14.

[27] Wandell BA, Dumoulin SO, Brewer AA. Visual field maps in human cortex. Neuron 2007;56:366–83. https://doi.org/10.1016/j.neuron.2007.10.012.

[28] Sereno AMI, Dale AM, Reppas JB, Kwong KK, Belliveau JW, Brady TJ, et al. Borders of Multiple Visual Areas in Humans Revealed by Functional Magnetic Resonance Imaging. Science 1995;268:889–93.

[29] Engel SA, Glover GH, Wandell BA. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. Cerebral Cortex 1997:181–92.

Supplement:

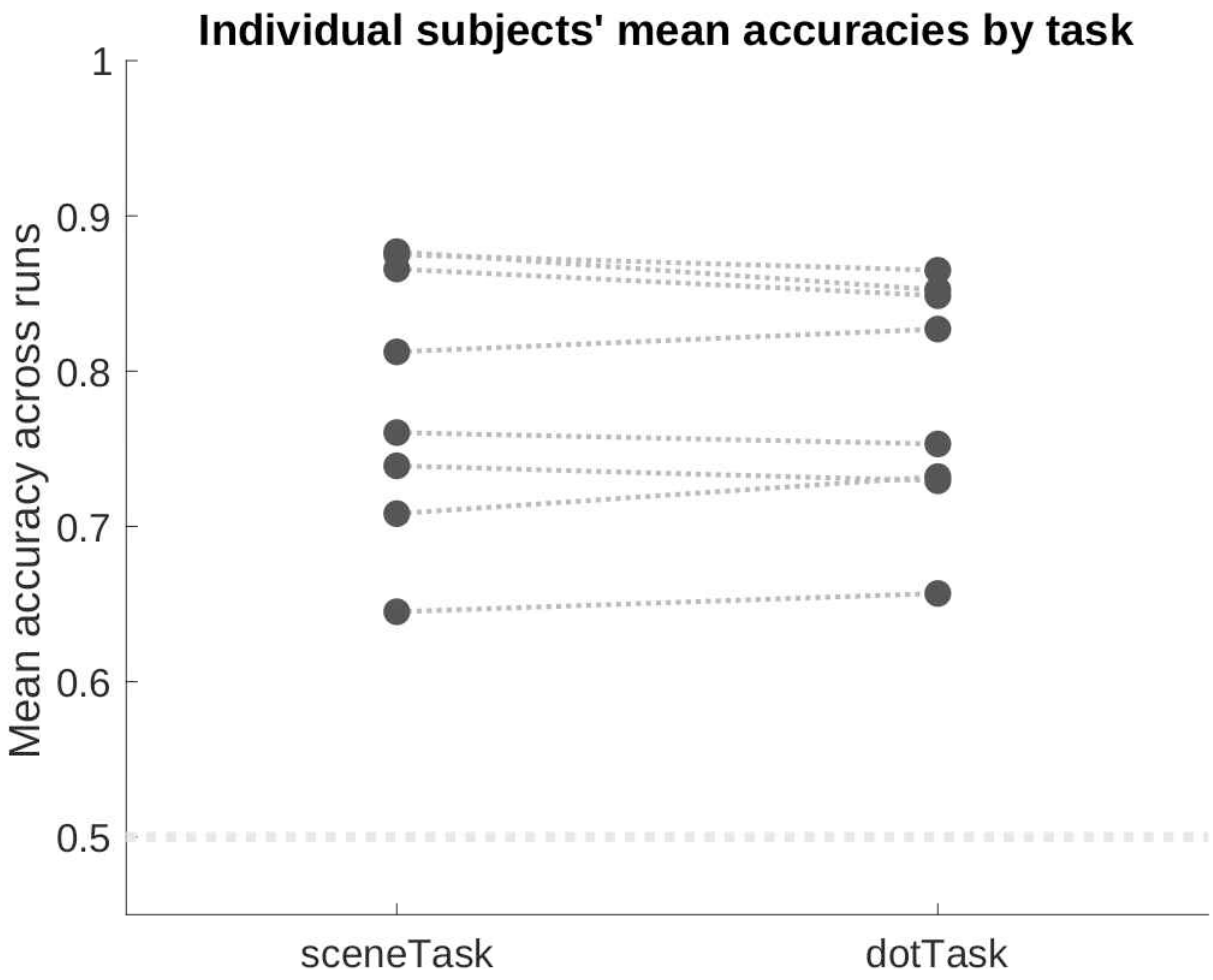


Figure S1. Participants' performance was well above chance, with no obvious outliers. The difficulty of the scene task was fixed because it depended on pre-determined placement of dots on the scene. We chose these placements with the goal of keeping participants off of ceiling. We were able to staircase performance in the dot task to match the scene task, avoiding large differences in performance across runs and enabling us to determine that participants were remaining alert.

Table S1. RDM-correlations between 2D gabor-wavelet model and 3D-global scene-surface models are reliable within stimulus sets across 3 distance metrics. For the present stimuli, only the 3D-global scene-surface model was included, as it most closely corresponded to Lescroart & Gallant's (2019) model.

Distance metric	**Present stimuli**	**Lescroart & Gallant (2019)**
Cosine distance	-0.002	0.274
Spearman's rho	-0.001	0.191
Pearson correlation	-0.003	0.277

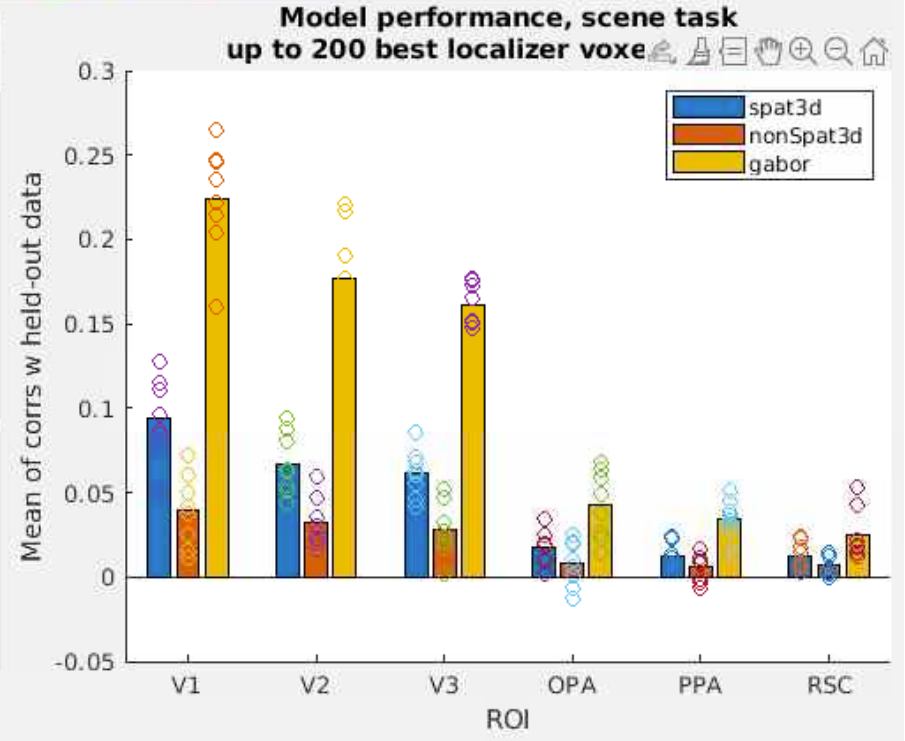


Figure S2: Scene-task-only performance. Magnitudes are almost identical to performance collapsed across task, suggesting that our lack of evidence for variance uniquely attributable to 3D surfaces is not due to our task manipulation.

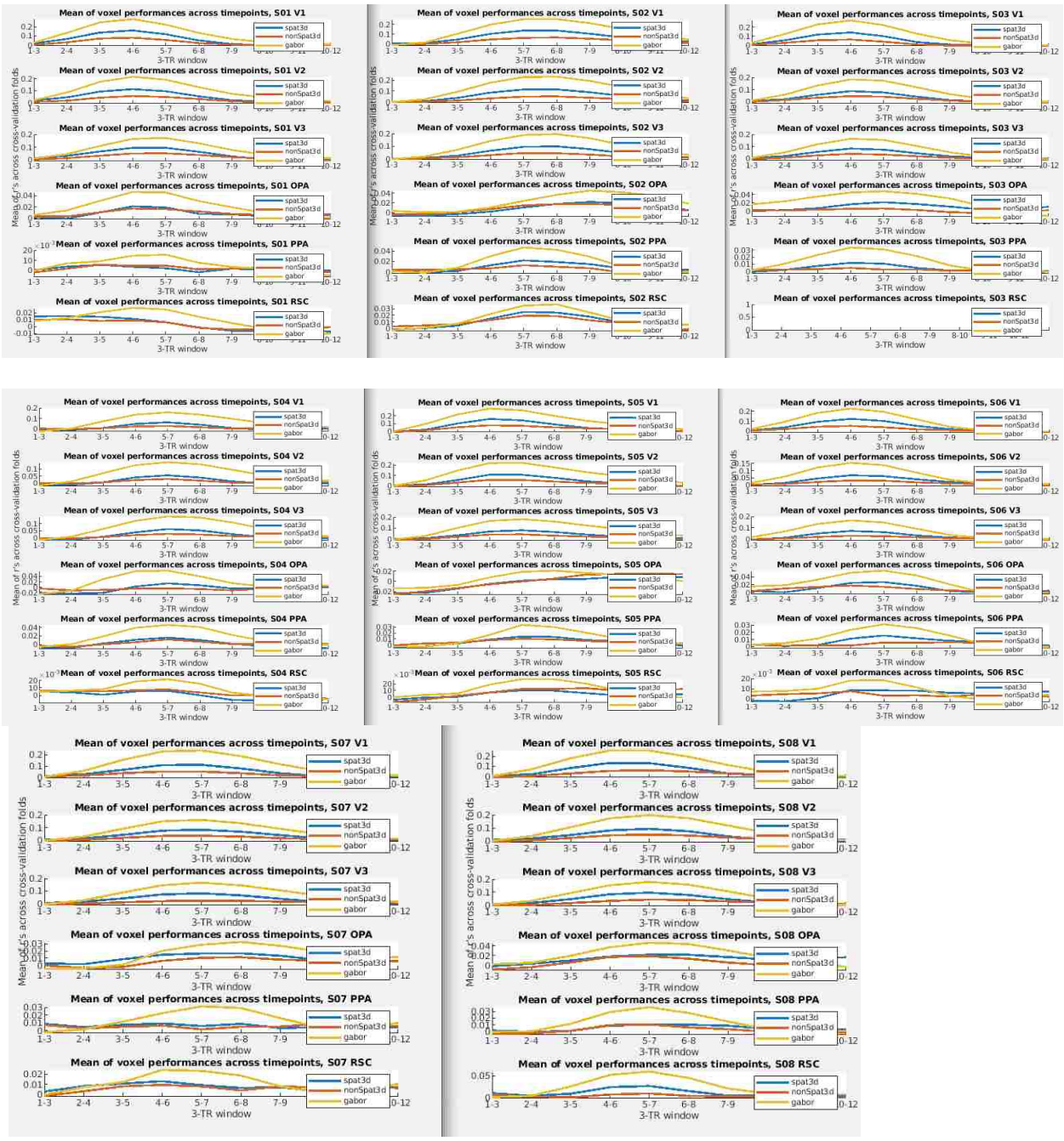
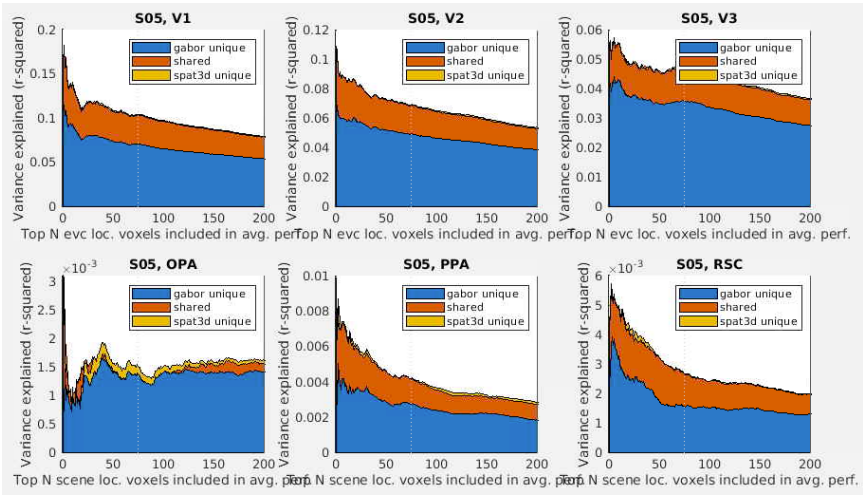


Figure S3 Sliding-window results. Prediction performance across ten 3-TR windows. The relative pattern of model performances is robust, even in individual subjects, arguing that our results are not due to differences in timepoints analyzed compared to other studies. The pane with no lines (S03, RSC) means that ROI combination did not have enough voxels (>=75) to be included in the analysis.



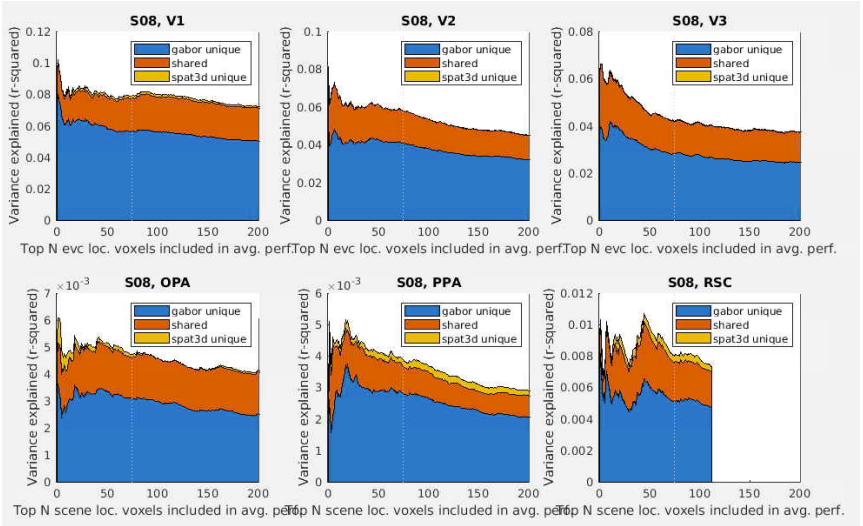


Figure S4: **Variance partitioning results are robust to ROI size.** Two example subjects' variance partitioning results plotted as a function of number of voxels included in each ROI. Voxels were ordered by independent-localizer activation: scrambled > baseline for EVC ROIs, top row; scenes > objects for scene-selective ROIs, bottom row. Then, results were plotted as though we'd defined 1-voxel ROIs, 2-voxel ROIs, etc. (See methods for details for ROI definition.) Plots are stacked, height-along-y-axis denoting: unique variance explained (r^2) by the gabor model (first legend entry; bottom layer of stacked plot; blue); shared variance (middle; red); and variance uniquely explained by the 3D model, in this case the 3D-quadrant-based model (bottom legend entry; top layer of stacked plot; yellow). Note that the other subjects' plots looked remarkably similar to these. The downward slopes of many plots indicate that our ROI sizes (200 in most cases, or the max number of available voxels above the significance cutoff—at least 75 required—for our independent localizer data) may have caused us to find a smaller magnitude of effects than in other studies that used a stricter cutoff.

Table S2A: Two-tailed p-values corresponding to permutation tests for each model compared to 0, as well as comparisons of performance for each.

	spat3d	nonSpat3d	gabor
V1	0.0001	0.0001	0.0001
V2	0.0002	0.0001	0.0001
V3	0.0001	0.0001	0.0001
OPA	0.0009	0.0132	0.0002
PPA	0.0016	0.0019	0.0001
RSC	0.0099	0.004	0.0018

	spat3dVsNonSpat3d	spat3dVsGabor	nonSpat3dVsGabor
V1	0.0001	0.0001	0.0001
V2	0.0002	0.0001	0.0001
V3	0.0001	0.0001	0.0001
OPA	0.0342	0.0002	0.0008
PPA	0.0221	0.0001	0.0001
RSC	0.1638	0.0005	0.0025

Table S2B: Variance partitioning, two-tailed p-values corresponding to permutation tests for 3D-quadrant-based ("spat3d") and gabor models.

	gabor unique	shared	spat3d unique
V1	0.0001	0.0002	0.0261
V2	0.0002	0.0002	0.0088
V3	0.0001	0.0001	0.1662
OPA	0.0006	0.0019	0.6623
PPA	0.0002	0.0004	0.4103
RSC	0.0257	0.0259	0.8987

	gaborUniqueVsShared	gaborUniqueVsSpat3dUnique	sharedVsSpat3dUnique
V1	0.0001	0.0001	0.0001
V2	0.0001	0.0002	0.0002
V3	0.0002	0.0001	0.0001
OPA	0.0027	0.0008	0.002
PPA	0.0011	0.0002	0.0002
RSC	0.0367	0.0155	0.0193

Table S2C Variance partitioning, two-tailed p-values corresponding to permutation tests for 3D-global ("nonSpat3d") and gabor models.

6×3 table

	gabor unique	shared	nonSpat3d unique
V1	0.0001	0.0007	0.0262
V2	0.0001	0.0001	0.7131
V3	0.0001	0.0003	0.3412
OPA	0.0018	0.0069	0.2607
PPA	0.0005	0.0018	0.1288
RSC	0.0526	0.0048	0.2874

```
>> varComparisonPvals

varComparisonPvals =
```

6×3 table

	gaborUniqueVsShared	gaborUniqueVsNonSpat3dUnique	sharedVsNonSpat3dUnique
V1	0.0001	0.0001	0.0007
V2	0.0002	0.0001	0.0001
V3	0.0001	0.0001	0.0002
OPA	0.0061	0.0016	0.004
PPA	0.0002	0.0003	0.0015
RSC	0.0793	0.0297	0.001