UNIVERSITY OF CALIFORNIA SAN DIEGO

Investigating 3D scene-surface information in memory and neural representations

A dissertation submitted in partial satisfaction
of the requirements for the degree Doctor of Philosophy

in

Experimental Psychology

by

Anna Shafer-Skelton

Committee in charge:

Professor Timothy Brady, Co-Chair
Professor John Serences, Co-Chair
Professor Eran Mukamel
Professor Douglas Nitz
Professor Ed Vul

2022

The Dissertation of Anna Shafer-Skelton is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I want to acknowledge the amazing support of my past and present advisors and mentors; members of the Brady, Serences, and Störmer labs for their support and friendship over the past 5.5 years; as well as friends and family who supported me, especially by sending me cute pet pictures.

I would also like to acknowledge generous funding from the National Science Foundation Graduate Research Fellowship program and the APA Dissertation Award.

Chapter 1, in full, is a reprint of the material as it appears in: Shafer-Skelton, A., & Brady, T. (2018). Scene layout priming relies primarily on low-level features rather than scene layout. The dissertation author was the primary investigator and author of this paper.

Chapters 2 and 3 will be prepared to submit for publication. Shafer-Skelton, A.; Brady, Timothy F.; Serences, John T. The dissertation was the primary investigator and author of this material.

VITA

2011	Bachelor of Arts, Washington University in St. Louis
2012	Bachelor of Fine Arts, Washington University in St. Louis
2013-2014	Research Assistant, Harvard Vision Lab
2014-2016	Lab Manager, OSU Vision & Cognitive Neuroscience Lab
2022	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Starks, M. D., Shafer-Skelton, A., Paradiso, M., Martinez, A. M., & Golomb, J. D. (2020). The influence of spatial location on same-different judgments of facial identity and expression. Journal of Experimental Psychology: Human Perception and Performance.

Shafer-Skelton, A., & Brady, T. (2019). Scene layout priming relies primarily on low-level features rather than scene layout. Journal of Vision, 19(1).

Brady, T. F., Störmer, V. S., Shafer-Skelton, A., Williams, J. R., Chapman, A. F., & Schill, H. M. (2019). Scaling up visual attention and visual working memory to the real world. In Psychology of Learning and motivation (Vol. 70, pp. 29-69). Academic Press.

Shafer-Skelton, A., & Golomb, J. D. (2018). Memory for retinotopic locations is more accurate than memory for spatiotopic locations, even for visually guided reaching. Psychonomic bulletin & review, 25(4), 1388-1398.

Bapat, A. N., Shafer-Skelton, A., Kupitz, C. N., & Golomb, J. D. (2017). Binding object features to locations: Does the "spatial congruency bias" update with object movement?. Attention, Perception, & Psychophysics, 79(6), 1682-1694.

Shafer-Skelton, A., Kupitz, C. N., & Golomb, J. D. (2017). Object-location binding across a saccade: A retinotopic spatial congruency bias. Attention, Perception, & Psychophysics, 79(3), 765-781.

Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. Journal of Experimental Psychology: Human Perception and Performance, 43(6), 1160.

ABSTRACT OF THE DISSERTATION

Investigating 3D scene-surface information in memory and neural representations

by

Anna Shafer-Skelton

Doctor of Philosophy in Experimental Psychology

University of California San Diego, 2022

Professor Timothy Brady, Co-Chair

Professor John Serences, Co-Chair

Visual attention and memory research focus heavily on controlled experiments with simple visual features and discrete objects, despite evidence for an important distinction between representations of objects and large extended surfaces in scenes. Other work suggests that what we know about attention and memory for objects may not simply extend to scene-surface information, motivating us to better characterize these differences. Chapter 1 investigates a paradigm widely cited as demonstrating the existence of scene-specific representations in working memory, finding that it does not convincingly differentiate scene-specific information in working memory from information in high-capacity iconic memory, which could be in a number of different formats. This led us to explore other approaches to more specifically isolate scene surface information. For example, recent fMRI work has revealed a potential marker for scene-specific representations, which showed promise for investigating influences of different task or attention conditions. In Chapter 2, I tested whether this scene-specific information persisted when participants viewed naturalistic scene photographs rather than 3D-rendered environments. I used deep neural networks to estimate ground-truth 3D information about stimuli in a publicly available fMRI data set. Using this information to predict fMRI responses, I found evidence of 3D scene-specific representations, although this information was less distinguishable from 2D information than in the previous work. In Chapter 3, I re-examined this finding using stimulus photographs with ground-truth 3D information that, as a set, had more potential to differentiate 3D-surface features from 2D features. We also tested the presence of more spatially precise scene-specific information that could be more useful in moving through the 3D world, finding a shocking dominance of 2D visual information over both types of 3D information, with no evidence for scene-specific representations. This echoes behavioral work suggesting that 2D textures may underlie 3D representations in natural scene images and highlights the importance of studying complex real-world information using complementary stimulus sets that preserve different aspects of the natural world. Together, these chapters lay critical groundwork for understanding how scene representations behave under different attention and memory conditions.

INTRODUCTION

Our visual systems encounter many types of visual information on a regular basis: simple features like color, orientation, and spatial frequency; discrete visual objects like an apple or a chair; and the large extended surfaces around us like walls and floors. A great deal of evidence across different disciplines argues that the distinction between discrete objects and large extended surfaces is important to our visual system—for example, human-developmental and rodent work suggests that they are prioritized differently over the course of development and across species (Cheng, 1986; Hermer & Spelke, 1994; Landau & Lakusta, 2009), human fMRI work shows these types of representations are anatomically distinct in the brain (e.g., Epstein & Kanwisher, 1998), and TMS work shows they can be selectively disrupted (e.g., Dilks, Julian, Paunov, & Kanwisher, 2013). Importantly, an intriguing set of work suggests that scene-specific and related types of information may be processed very rapidly or with reduced focal attention (Alvarez & Oliva, 2009; Greene & Oliva, 2009), suggesting that we shouldn't assume that what we know about attention and memory for visual objects and simple features extends to scene information. This motivated our use of a paradigm widely cited as evidence for abstract scene-layout information in memory (Chapter 1), as well as a promising new technique to isolate scene-specific visual information in scene-selective cortical areas (Chapters 2 and 3).

Chapter 1: Lack of knowledge about scene surface information in visual working memory

Visual working memory is a particularly critical realm of visual processing to understand, since it has the potential aid in constructing our mental model of the world through (1) maintaining external visual "landmarks" that we encode other objects in relation to (possibly aiding in the integration of visual information across eye movements, e.g.), and (2) allowing us to access information that is currently not foveated or is out of view. While our severely limited working memory capacity appears to be an obstacle to these processes, I argue in Brady et al. (2019) that this limited capacity is an artifact of how we test working memory—while the majority of working memory paradigms measure capacity in terms of how many colors (or other abstract objects) someone can remember, the real world contains many structural regularities and more diverse types of visual information, such as the large extended surfaces that make up a scene. Is such scene information stored in working memory separately from objects? If so, it is possible that memory stores that are at least to some degree domain-specific could effectively expand our working memory capacity, facilitating less-overlapping memory representations that are less likely to compete for resources.

To look into this possibility, Chapter 1 of this dissertation investigates a paradigm widely cited as demonstrating the maintenance of abstract scene layout information in memory. We found evidence that this established paradigm doesn't actually isolate scene layout information (and may instead be driven by a more fragile form of memory, which, unlike visual working memory, is thought to be disrupted by eye movements). Because this previous paradigm hinged on the assumption that effects were driven by working-memory representations, an assumption that significantly narrowed the hypothesis space about the types of visual information that could be contributing, our findings argue that this work does not isolate scene-surface information. This highlights a giant gap in the existing literature: we know very little about the how and under which circumstances the visual system stores representations of the 3D geometry of a visual space, motivating Chapters 2 and 3 of my dissertation.

Chapter 2: Deep-net-derived surface estimations of natural scenes predict voxel responses in scene-selective cortex

It is difficult to design an experiment well-controlled enough to isolate scene-surface information (or any other type of more complex visual information), especially since lower-level information might actually underlie it (Brady, Shafer-Skelton, & Alvarez, 2017) and so may be impossible to disentangle. Here, we look to recent neuroimaging approaches that seek to quantify the amount of unique or overlapping variance that different types of visual information can explain in human fMRI responses to visual stimuli (e.g., Groen et al., 2017; Lescroart & Gallant, 2019). Using this approach, Lescroart & Gallant (2019) found evidence for information that is a likely candidate for a marker of scene-specific information in the brain. Their stimuli were computer-rendered 3D images that had ground-truth 3D depth information, and they generated features corresponding to those images that summarized the 3D spatial characteristics of surfaces across each image. They found that a portion of voxel responses in scene-selective areas OPA, PPA, and RSC could be uniquely attributed to these features, an indication of a scene-specific representation.

While Lescroart & Gallant (2019) use 3D-rendered images, we believed it was important to ensure that their findings extended to scene photographs, which contain more naturalistic texture information that scene-selective cortex may be sensitive to. We use a publicly available fMRI data set (Chang et al., 2019) and generated estimated 3D distance and surface-direction information using pre-trained DNNs (Zamir et al., 2019). Using this information, we find a similar pattern of results as Lescroart & Gallant (2019), although with a smaller magnitude of voxel responses uniquely attributable to 3D scene surface information and the largest amount shared between 2D and 3D features. This pattern converges with behavioral evidence that patterns of 2D and 3D orientation may underlie our scene processing abilities (Brady, Shafer-Skelton, & Alvarez, 2017).

Chapter 3: Stimulus dependence of 3D-scene-surface representations in scene-selective cortex

In Chapter 3, we collected our own data, seeking to (1) assess the presence of more navigationally relevant scene-surface information, and (2) further explore the relationship between 2D and 3D visual features. Starting from a large collection of natural scene images (~4 million) with ground-truth distance and surface-direction annotations (used to train the DNNs used in Chapter 2), we selected images so that, across images, global summaries of 3D information (Ch. 2; Lescroart & Gallant, 2019) covaried minimally with spatially specific (quadrant-based) summaries of 3D information that might be more useful for navigating. In doing so, we found that there was also a smaller relationship between 2D and 3D visual features than in previous work, setting us up for a stronger test of the presence of scene-specific information (via voxel responses uniquely attributable to a 3D scene-surface model).

We were surprised that, using this new stimulus set, we now found no evidence for scene-specific representations—while 3D model performance was significantly above 0 in scene-selective areas, our variance partitioning analysis found that this must have been due to the influence of shared 2D/3D information and not information uniquely attributable to 3D scene-surface representations. Follow-up analyses argue that our results cannot be straightforwardly explained by analysis differences. Instead, one possibility is that this difference can be explained by stimulus differences—previous work used artificially generated stimulus sets with less naturalistic texture information than photographs. When we used naturalistic scene photographs, this may have encouraged our participants' visual systems to use the same (2D) cues for scene processing that they can rely on in the real world. Another intriguing possibility is that, while some aspects of Lescroart & Gallant's (2019) stimuli were less naturalistic, the fact that they were movies means they did contain another depth cue that ours did not: motion parallax. Thus our results also argue that it is important to test the impacts on scene-surface information when motion-parallax can vs. cannot be used as a cue.

These chapters are motivated by an important distinction between discrete-object representations and scene-surface representations. We investigate techniques with the potential to study scene-surface representations in a behavioral memory paradigm (Chapter 1), as well during attentional/task manipulations in fMRI (Chapters 2 and 3). Together, these projects make important strides towards understanding the separability of scene-specific representations from other types of information, as well as more precisely understanding the format of the information represented in scene-selective cortex. Finally, they lay critical groundwork for understanding how scene-specific representations are affected under different attentional and memory conditions.

References

Alvarez, G. a, & Oliva, A. (2009). 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, 106(18), 7345–7350. https://doi.org/10.1073/pnas.0808981106

Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. Journal of Experimental Psychology: Human Perception and Performance, 5, 0–17. https://doi.org/10.1037/xhp0000399

Brady, T. F., Störmer, V. S., Shafer-Skelton, A., Williams, J. R., Chapman, A. F., & Schill, H. M. (2019). Scaling up visual attention and visual working memory to the real world. Psychology of Learning and Motivation - Advances in Research and Theory, 70, 29–69. https://doi.org/10.1016/bs.plm.2019.03.001

Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., & Aminoff, E. M. (2019). BOLD5000, a public fMRI dataset while viewing 5000 visual images. Scientific Data, 6(1), 49. https://doi.org/10.1038/s41597-019-0052-3

Cheng, K. (1986). A purely geometric module in the rat's spatial representation*. Cognition, 23, 149–178.

Dilks, D. D., Julian, J. B., Paunov, A. M., & Kanwisher, N. (2013). The occipital place area is causally and selectively involved in scene perception. Journal of Neuroscience, 33(4), 1331–1336. https://doi.org/10.1523/JNEUROSCI.4081-12.2013

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. Nature, 392(6676), 598–601. https://doi.org/10.1038/33402

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: seeing the forest without representing the trees. Cognitive Psychology, 58(2), 137–176. https://doi.org/10.1016/j.cogpsych.2008.06.001

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2017). 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, 207530. https://doi.org/10.1101/207530

Hermer, L., & Spelke, E. S. (1994). A geometric process for spatial reorientation in young children. Nature, 370(6484), 57–59. https://doi.org/10.1038/370057a0

Landau, B., & Lakusta, L. (2009). Spatial representation across species: geometry, language, and maps. Current Opinion in Neurobiology, 19(1), 12–19. https://doi.org/10.1016/j.conb.2009.02.001

Lescroart, M. D., & Gallant, J. L. (2019). Human Scene-Selective Areas Represent 3D Configurations of Surfaces. Neuron, 101(1), 178–192.

https://doi.org/10.1016/j.neuron.2018.11.004

Zamir, A., Sax, A., Shen, W., Guibas, L., Malik, J., & Savarese, S. (2019). Taskonomy: Disentangling task transfer learning. In IJCAI International Joint Conference on Artificial Intelligence (Vol. 2019-Augus, pp. 6241–6245). https://doi.org/10.24963/ijcai.2019/871

CHAPTER 1: Scene layout priming relies primarily on low-level features rather than scene layout.

Abstract

The ability to perceive and remember the spatial layout of a scene is critical to understanding the visual world, both for navigation and for other complex tasks that depend upon the structure of the current environment. However, surprisingly little work has investigated how and when scene layout information is maintained in memory. One prominent line of work investigating this issue is a scene priming paradigm (e.g., Sanocki & Epstein, 1997), in which different types of previews are presented to participants shortly before they judge which of two regions of a scene is closer in depth to the viewer. Experiments using this paradigm have been widely cited as evidence that scene layout information is stored across brief delays and have been used to investigate the structure of the representations underlying memory for scene layout. In the present experiments, we better characterize these scene priming effects. We find that a large amount of visual detail rather than the presence of depth information is necessary for the priming effect; that participants show a preview benefit for a judgment completely unrelated to the scene itself; and that preview benefits are susceptible to masking and quickly decay. Together, these results suggest that "scene priming" effects do not isolate scene layout information in memory, and that they may arise from low-level visual information held in sensory memory. This broadens the range of interpretations of scene priming effects and suggests that other paradigms may need to be developed to selectively investigate how we represent scene layout information in memory.

Introduction

One of the central challenges in understanding our visual experience is understanding what information about the world we hold in visual memory across brief delays and interruptions, like eye movements and blinks. Visual memory is critical for many tasks we perform every day, like visual search and spatial navigation, and given our limited ability to process everything from a single fixation, visual memory is necessary to build up an experience of a coherent and complete visual scene (e.g., Hollingworth, 2004, 2005). Countless studies investigate memory for discrete *objects*, including the capacity limit of visual memory for objects (e.g., Brady et al., 2016), the format of the representations for objects and how precision and the number of objects held in mind trade-off (Zhang & Luck, 2008), and what neural mechanisms are responsible for storing objects in working memory (Serences, 2016).

However, our visual environment is made up both of discrete objects and also of extended surfaces which form a spatial layout, and there is significant evidence that our visual system processes these types of information separately. For example, fMRI studies in humans show evidence for regions of the brain that respond selectively to scenes compared to objects (Epstein, 2005; Epstein & Kanwisher, 1998; Kravitz, Saleem, Baker, & Mishkin, 2011) and which seem to represent features of a scene's spatial layout rather than the objects it contains (Epstein, 2005; Park, Brady, Greene, & Oliva, 2011). In addition, it is possible to recognize briefly presented scenes even without being able to recognize any of the objects in those scenes (Oliva & Torralba, 2001; Schyns & Oliva, 1994), providing evidence of the independence of scene recognition from object recognition. Greene & Oliva (2009) proposed that this ability could arise from the representation of global properties of scenes, such as the "perspective" or "openness" of a scene. Past research has also drawn distinctions between other types of scene information that may be represented, for example: scene meaning (sometimes called “gist”; e.g., if the scene is a beach, a dining room, etc.) (Oliva, 2005) and the spatial layout of scenes (Epstein, 2005). Finally, evidence suggests that scene structure, including the spatial layout of a scene, is crucial to guiding our attention during visual search for objects, and may be represented in a global way independent of object processing (e.g., Torralba, Oliva, Castelhano, & Henderson, 2006; Wolfe, Võ, Evans, & Greene, 2011). However, despite this evidence for distinct representations of scenes (separate from those of objects), little work has investigated how scene-specific spatial layout information is maintained across saccades or brief delays, with most work on scene memory focusing on the role of memory for objects within scenes (Hollingworth, 2004, 2005).

One technique used to study memory for natural scenes in general is to test whether a preview of a scene facilitates subsequent processing related to that scene. For example, a preview of a real-world scene image facilitates subsequent visual search for an object present in that scene (Castelhano & Henderson, 2007; Võ & Henderson, 2010). While there is evidence that the memory representations retained in these studies are abstracted from the exact visual features (e.g., Castelhano & Henderson, 2007 show size invariance), these studies do not make it clear what specifically about the scene is remembered across the delay or to what extent this memory reflects the spatial layout per se as opposed to hypotheses about particular objects and their locations. Work by Sanocki and colleagues has asked more directly about the extent to which the spatial layout of a scene is held in memory by examining the conditions under which a preview of a scene facilitates a depth judgment within that scene (e.g., Sanocki, 2003, 2013; Sanocki & Epstein, 1997; Sanocki, Michelet, Sellers, & Reynolds, 2006). Deciding which of two things is closer in depth specifically targets scene layout representation as it requires participants to have processed and held in mind information about which parts of a scene are near or far from the observer, as opposed to only having held in mind a distribution of possible locations of objects. This "scene priming" paradigm is widely cited as an example of scene layout information being maintained in memory (e.g., by Chun & Jiang, 1998; Oliva & Torralba, 2001). However, while existing experiments show that the effect persists when some low-level information is varied (e.g., Sanocki, 2003), the effect is often diminished, and it remains possible that low-level visual information (e.g., patterns of orientation across the image; e.g., Brady, Shafer-Skelton, & Alvarez, 2017) could be driving the effect without an abstract representation of the spatial layout of a scene.

In the present experiments, we sought to better characterize the robustness and content of the memory representations responsible for scene priming effects. In particular, we ask (1) whether scene priming paradigms are able to isolate the effects of scene layout information held in memory, and (2) whether scene priming effects are primarily driven by information held in maskable memory stores, such as iconic memory, or more robust memory stores, such as visual working memory. In our first experiment, we reasoned that if "scene priming" benefits reflect memory for scene layout, we would expect them to persist when scene previews contain layout information (boundaries of major surfaces or large objects), even if these previews have no identifiable objects and little extraneous visual detail. However, in Experiment 1 we find that while previews consisting of full photographs of target scenes are able to speed depth judgments on the target scenes, sparse line drawings of the scenes, which contain only the boundaries of major surfaces or objects and lack semantic information, are unable to speed depth judgments despite containing significant depth information. In Experiment 2 we find that even in a task that doesn't require the usage of the scene at all — and particularly not its layout — photo preview benefits are still present, suggesting they are not a selective index of scene layout or even scene processing. In Experiment 3, we test whether scene priming benefits are due to a memory store robust to visual masking (e.g., working memory). We find a preview effect for the more detailed line drawings used by Sanocki and Epstein (1997), which contain identifiable shapes as well as extra visual detail, and we find that it is abolished with a mask and a longer delay. This suggests that even line drawing preview benefits may be due to a maskable memory store, such as iconic memory. Compared to previous interpretations, these results broaden the possibilities for how the preview is speeding participants' judgments—arguing that low-level information held in iconic memory may be sufficient to facilitate the detection of sudden onsets of the target shapes rather than giving participants a head start on processing scene layout.

Experiment 1: Preview benefit for photos but not sparse line drawings

In a first experiment we tested whether participants were faster at making a depth judgment (i.e., which of two regions of a scene would be closer in depth) when they first saw a preview of either a photograph of the scene or a line drawing of the scene, as compared to an uninformative rectangle presented with the same timing as the two scene-specific previews. The main task for participants was to judge which of two red dots on a scene was on the position in the scene that was closer in depth to the viewer (Figure 1; see Sanocki, 2003). Just before each scene was presented, participants saw one of the preview images. Because line drawings share minimal low-level visual features with the target images, a line drawing preview benefit might indicate that scene priming effects are due to abstract information stored in memory about the spatial layout of the surfaces in the scene. To best assess this, the line drawings we selected for this experiment contained the boundaries of the major surfaces and objects in a scene but were screened to ensure they contained no recognizable objects. Because they were automatically generated from the boundaries dividing labeled regions of a scene, they also did not contain extraneous visual detail (e.g., blades of grass, artistic details).

Experiment 1 Method

The design, number of participants, and analysis plan for this experiment were preregistered (URL for this experiment: https://aspredicted.org/yw5bg.pdf; see Supplemental Materials for all pre-registrations).

Participants: To complete a full counterbalance (see Design & procedure for details), we had 102 participants (6 groups of 17 each). Participants were Mechanical Turk workers who participated in exchange for monetary compensation. Previous literature finds that Mechanical Turk workers are representative of the adult American population (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011) and provide similar data to participants run in laboratory visual cognition studies (Brady & Alvarez, 2011). We recorded timing information in order to ensure consistency across individual participants' computers and monitors.

Stimuli: Fifty-four images of indoor scenes were selected from the SUNRGB-D database (Song, Lichtenberg, & Xiao, 2015), which includes RGB images of scenes as well as corresponding semantic segmentations and maps of ground-truth depth. Because we didn’t want participants to be able to use the vertical position of the target dots as a depth cue, the two target dots placed on each image always had the same vertical position and different horizontal position in the image. Left-right depth-asymmetric scenes ensured a wider variety of possible target dot locations. Thus, to select the scenes to use as target images, we first ordered the images by asymmetry in the mean depth between the left and right halves of the image. Starting with the most depth-asymmetric scenes, line drawings were then created in Matlab by tracing the borders of the semantic segmentations of these same images, and the first ~500 line drawings were screened for identifiable objects, as we wished our line drawing preview images to contain information about spatial layout but not about the identity of particular objects. Participants were asked to list any objects they could identify in the images (excluding major surfaces, like “wall” or “floor”), and an image was selected for the main experiments if neither author AS nor any of 10 pilot participants per image reported being able to identify any objects. This resulted in 54 images. One set of probe locations was chosen for each image, and target images were created by using Matlab to add red dots with white outlines at the chosen probe locations. Matlab was also used to create the rectangle preview. Scene photograph previews were the original scene images used to create target images. All images were cropped and down-sized, if necessary, to 561 x 427 pixels.

Design & procedure: Participants’ task on every trial was to judge which of two red probe dots was on the part of the scene image that would be closer to the viewer in depth in real life. Each trial began with a preview from one of three conditions: (1) a line drawing of the scene photo (line drawing preview); (2) the black outline of a rectangle (rectangle preview), as used in Sanocki & Epstein (1997); and (3) the exact same scene photograph that was used to create the target image (photo preview). Each preview image was presented for one second. Following a brief blank (87 ms, as in Sanocki & Epstein, 1997), the target image was presented until participants responded (see Figure 1 for a schematic of a trial). Participants were instructed to respond as quickly as possible while still getting most trials correct, and feedback was given for incorrect answers.

Each image appeared once in each of the three conditions. The order images appeared in was randomized with the constraint that each target image was presented for the first time before any images were presented for the second time. Six possible counterbalance conditions ensured that across all participants, each image appeared equally often in each of the six possible orders of preview conditions (e.g., line drawing, then photo, then rectangle; etc.).

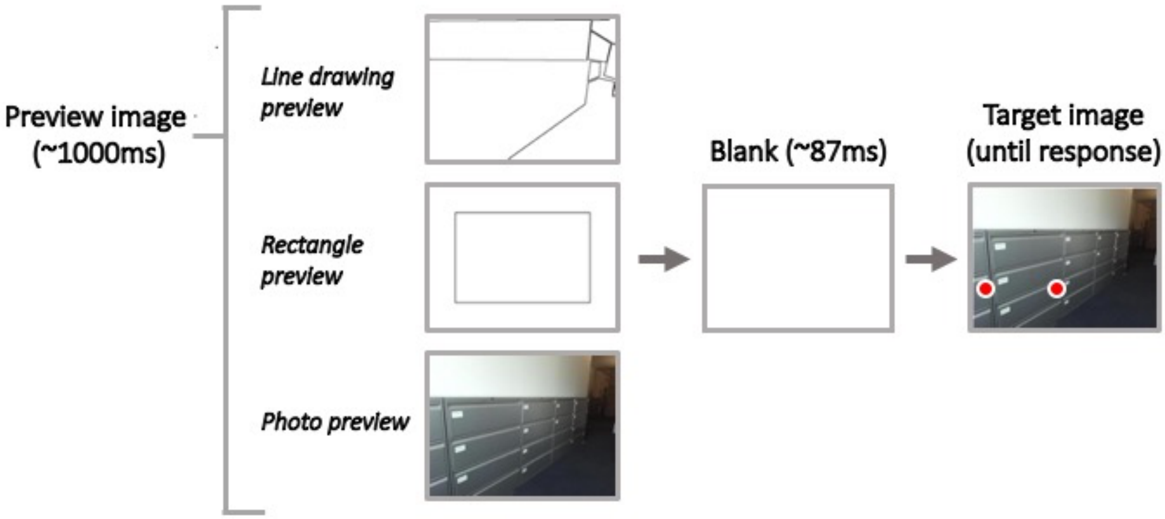


Figure 1.1: Trial timing and conditions for Experiment 1. Each trial started with a preview image from one of the three preview conditions -- a photo preview without the red probe dots present, a rectangle preview, or a line drawing preview that contained information about the spatial layout of the scene but not about the identity of individual objects. As in previous work, these previews were visible for 1000ms. After an 87ms blank, a target image was then presented, and participants were instructed to respond which of the locations cued by the two red probe dots would be closer to the viewer in depth in real life. (Red dots enlarged here for visibility.) In Experiment 1, preview conditions were intermixed, and participants were given no special instructions regarding the previews.

Analyses: Our exclusion criteria and analyses were decided in advance (see pre-registration). We excluded individual trials if reaction times were faster than 150ms and only included correct trials in reaction time analyses. Participants were excluded and replaced with a new participant from the same counterbalance condition if any of the following applied: overall accuracy more than 3 standard deviations below the mean accuracy; overall accuracy below 55%; same response key used on more than 80% of trials; median RT slower than 2 seconds for any of the three preview conditions; fewer than 50% of trials included in the main analysis, either because of RTs below 150ms, or because of incorrect responses. These criteria resulted in the exclusion of 15 participants (14 participants for accuracy, one of whom also had too many RTs faster than 150ms and another of whom also had median RTs slower than 2 seconds; as well as 1 participant for having median RTs slower than 2 seconds).

In all experiments, our statistics were performed based on each participants' median reaction time in each of the three preview conditions. The critical analyses were two t-tests between participants' median RTs in the photo preview condition and the rectangle preview condition, and between the line drawing preview condition and the rectangle preview condition. Effect sizes were calculated using Cohen's *d*.

Experiment 1 Results

Participants were faster with photo previews (M=857 ms) than with rectangle previews (M=900 ms; t(101) = 4.91, p < 0.001, *d* = 0.49), indicating that participants were making use of the previews. However, we did not see facilitation for the line drawing preview condition (M=900 ms) compared to the rectangle preview condition (M=900 ms; t(101) = -0.07, p = 0.94, *d* = -0.06). The photo preview benefit was also significantly larger than the line drawing preview benefit (t(101) = 5.64, $p < 0.001$, d = 0.56).

Because we designed the task to have as many usable trials as possible for the reaction time analysis, mean accuracies were high and within a 0.7% range (line drawing: 97.5%, rectangle: 97.0%, photo: 96.8%). Uncorrected post-hoc t-tests showed one significant accuracy difference (line drawing vs. photo) and small effect sizes in each comparison (rect vs. photo: t(101) = 0.49, p = 0.62, d = 0.05; line drawing vs. rect: t(101) = -1.92, p = 0.06, d = -0.19; line drawing vs. photo: t(101) = -2.32, p = 0.02, d = -0.23). Because there are no large accuracy differences, speed-accuracy tradeoffs are unlikely to have affected our pattern of RT data. See Figures A4-A6 for accuracy data, including individual subject accuracies.

To verify that our line drawings contained information about the spatial structure of each scene, we performed a supplemental experiment (see Experiment A1), in which the red target dot locations were placed directly on the line drawings, and participants judged which regions of the line drawings would be closer in real life. Participants saw the line drawings for the same timing as they saw them during the preview in Exp. 1 (1000ms). Participants were 67% accurate at this task, significantly above chance (t(99) = 17.46, $p < 0.001$, *d* = 1.75), and in a post-hoc analysis, when we re-analyzed Experiment 1 using only the line drawings with significantly above-chance performance (lowest: 66%; mean: 78%), we again did not find a line drawing preview benefit (t(101) = 0.21, p = 0.83, d = 0.02). Again, the photograph preview benefit and the interaction between the line drawing and photograph preview benefits were both significant (photo preview benefit: t(101) = 4.98, $p < 0.001$, d = 0.49; interaction: t(101) = 5.17, $p < 0.001$, d = 0.51). In order to further explore the relationship between depth information in the sparse line drawing previews and the line drawing preview benefit, we also plotted the size of the line drawing preview benefit for each image against the proportion of participants who correctly judged depth in that image. If our lack of a preview benefit were due to lack of depth information in the previews, we would expect a positive relationship between depth judgment accuracy and line drawing preview benefits. Instead, we find no evidence of a relationship ($r = 0.13$, $p = 0.35$; see Figure 3 for plot).

Experiment 1 Discussion

We found that while previews of the full photograph provided a significant benefit in a subsequent depth judgment task, sparse line drawing previews did not provide a benefit (relative to uninformative rectangle previews). This was true despite the presence of significant depth information in the line drawing previews and held even when we limited our analysis to only those line drawings that provided the best depth information.

In additional experiments reported in the Appendix, we replicated the photograph preview benefit (Experiments A1-A3) and the lack of a line drawing benefit (Experiments A1-A2; no line drawings were included in Experiment A3). These replications were originally designed to address the role of mirroring the photo or line drawing preview to distinguish representations of spatial layout from more global scene properties. In all experiments conducted using our sparse line drawing stimuli, we found the same pattern of results: a significant preview benefit for the photo previews, but none for the sparse line drawings in any of the 3 experiments in which they were included. This was despite the fact that these line drawings contain enough information for participants to make depth judgments.

Thus, despite the presence of depth information in our sparse line drawings, they did not lead to a preview benefit. Previous work (e.g., Sanocki & Epstein, 1997) has found reliable preview benefits from a different set of line drawings, an effect we successfully replicate in Experiment 3. There are two important differences between these stimulus sets. First, while Sanocki & Epstein's original (1997) drawings contained semantic information, we specifically chose line drawings that did not contain identifiable objects. This was because we wanted to be able to differentiate between effects due to the presence of semantic information vs. the presence of spatial layout. The second difference is that the original line drawings share much more local orientation information with the target images (e.g., from blades of grass, small and medium-sized objects) than the sparse line drawings used in Experiment 1. Critically, Experiment 3 of Sanocki & Epstein (1997) does show a scene priming benefit for artificially generated stimuli that lack semantic information (as our line drawings do) but also share much of the same local orientation information with the target images (which our line drawings do not). This led us to believe that the lack of a line drawing benefit in Experiment 1 was not due to the lack of semantic information or participants' inability to categorize our line drawings—instead, one important possibility to consider was whether the amount of visual detail (e.g., orientation information) shared between the previews and targets is critical to finding a line drawing preview effect, and that such a preview effect might not result from processing of scene layout.

Given the very brief delay in our experiment (87 ms, based on previous scene priming paradigms), it is possible that low-level visual information about the preview image may be stored in a high-capacity visual memory store, such as iconic memory, and that a preview image that is sufficiently similar to the target image (simply missing the probe dots) might allow participants to find the probe dots more efficiently. In other words, rather than giving participants a head-start on layout processing, it is also possible that when more visual detail is shared between the preview image and the target image, the sudden onset of the probe dots becomes more salient, speeding participants' judgments by speeding their detection of the probe dots (e.g., Jonides & Yantis, 1988; Theeuwes, 1991). To address this, we conducted two further experiments. Experiment 2 tests whether the photo preview benefit remains for a task in which participants' judgments on the target image should not be sped by knowledge of scene layout, as the target scene is irrelevant to the task, but could be sped by faster detection of the probe dots. Experiment 3 tests whether previews with more detailed line drawings facilitate depth judgements and tests how robust this is to longer delays and visual masking.

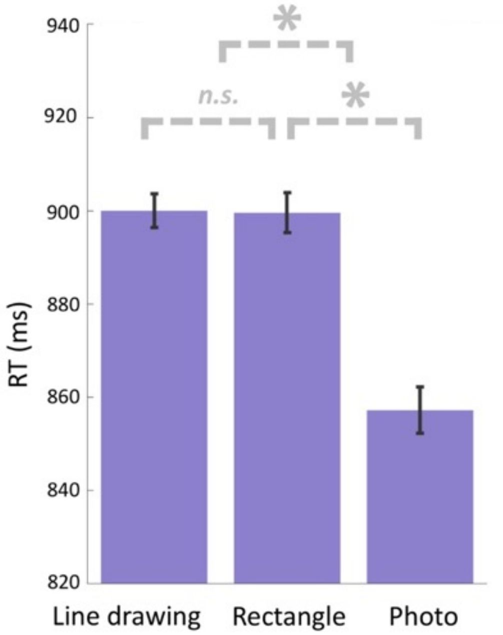


Figure 1.2: Participants' reaction times in each preview condition in Experiment 1. Bars represent means over participants. Error bars are within-participant SEM. N = 102.

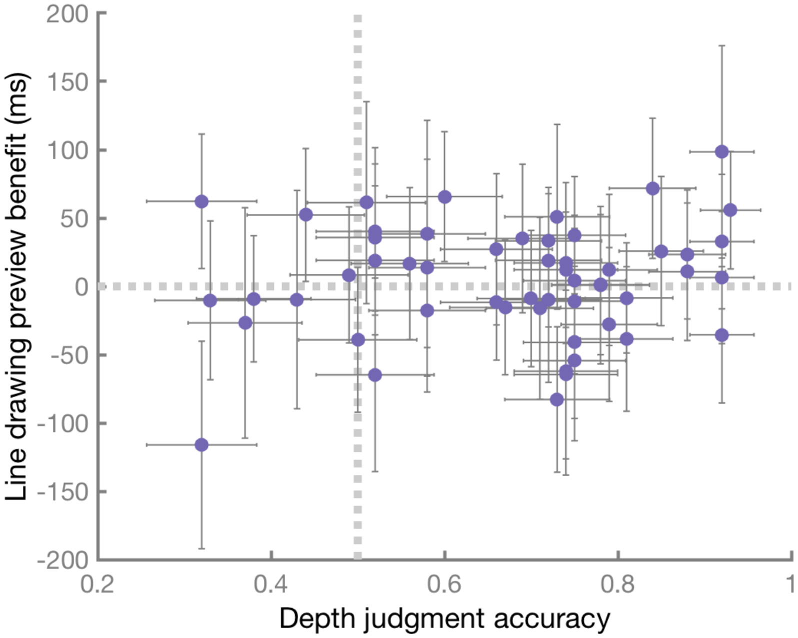


Figure 1.3: For each image, proportion of participants who correctly made the depth judgment in Experiment A1, plotted against the size of the line drawing preview benefit for that image in Experiment 1. Error bars on depth judgment accuracy are standard error of the proportion, and error bars on the line drawing preview benefit are SEM. Gray dotted lines indicate a line drawing preview benefit of 0 (horizontal) and chance performance on the depth judgment task (vertical).

Experiment 2: Photo preview benefit even when layout information is irrelevant

The sudden onset of an object tends to draw attention (Jonides & Yantis, 1988; Theeuwes, 1991), and thus the appearance of probe dots may draw attention even when the preview scene is in iconic memory rather than present on the screen. For example, empty-cell localization tasks and other related tasks show evidence for integration – and detection of new information – across brief delays (Di Lollo, 1980; Eriksen & Collins, 1967).

In particular, evidence suggests that if the delay between two stimuli is less than 80–100 milliseconds, visual persistence of the first overlaps with the initial sensory processing of the second, allowing participants to perceptually combine the two stimuli (Di Lollo, 1980; Eriksen & Collins, 1967), as in the case of two sets of dots forming a letter string (Eriksen & Collins, 1967).

Even at slightly longer delays, participants may be able to use informational persistence in iconic memory to notice the sudden onset of the probe dots (e.g., Hollingworth, Hyun, & Zhang, 2005). Thus, given the short delay used in typical scene priming experiments, it may be that much of the scene priming benefit arises as a result of faster detection of the probe items following the informative previews rather than faster processing of the target scene.

If preview benefits for more visually detailed preview images are driven by something other than scene layout information (e.g., speedier detection of the probe dots when more visual detail is shared between the preview and target images), we should find a preview benefit for a task that does not require scene layout information at all, or even the use of the target scene at all.

Thus, in Experiment 2, we used the same scene images and target shape locations as Experiment 1, but rather than seeing two red circles and making a depth judgment about the scene regions underlying these two circles, participants saw a red square and a red diamond and judged whether the left or right of these two target shapes was a square—a judgment for which the background scene was completely irrelevant. If participants' responses in scene priming experiments like Experiment 1 were speeded due to ease in locating the target shapes, we should also find a photo preview benefit here. On the other hand, if the scene priming paradigm effectively isolates a head-start in processing layout information, we should not expect a photo preview benefit, since layout information and scene information in general is not informative for this task.

Experiment 2 Method

The design, set size, and analysis plan for this experiment were preregistered (https://aspredicted.org/8g5v2.pdf; see Supplemental Materials for all pre-registrations).

Participants: Participants were 100 Mechanical Turk workers (25 in each of 4 counterbalance conditions) who participated in exchange for monetary compensation. No participants participated in the previous experiment.

Stimuli: Stimuli were the same as Experiment 1, except (1) we did not include a line drawing condition, since we did not find a line drawing preview benefit in Experiment 1, and (2) we replaced each set of the target dots with a square and a diamond.

Design and procedure: See Figure 4 for example trial. The design of this experiment was the same as for Experiment 1, except that there was no line drawing preview condition. This resulted in 4 counterbalance groups, since each target image was repeated with the opposite placement of squares and diamonds across groups, and each variation of each target image was presented either in the rectangle condition first or in the photo condition first across groups. Rectangle and photo previews were intermixed.

Participants' task was to judge whether the square was the left of the two shapes or the right of the two shapes.

Analyses: Analyses were the same as in Experiment 1, and exclusion criteria were the same as in the other 2 experiments. The preregistered exclusion criteria resulted in the exclusion of one participant for having an overall accuracy lower than three standard deviations below the mean accuracy. This participant was replaced with a participant from the same counterbalance condition.

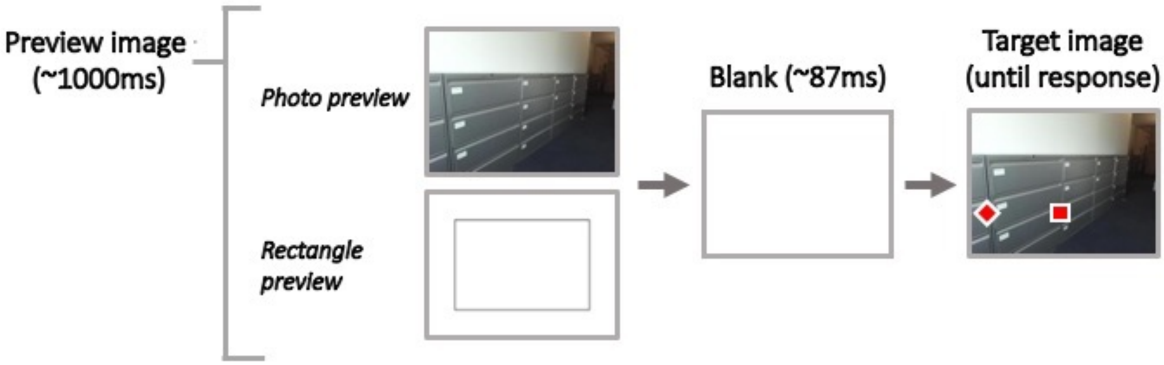


Figure 1.4: Trial timing and conditions for Experiment 2. As in Experiment 1, a preview image appeared for 1000ms, followed by an 87ms blank. In this experiment, each preview image was either the photo preview (without the square/diamond) or an uninformative rectangle preview. After the delay, a target image was presented, and participants were instructed to indicate which of the two shapes was a square (left or right). Square and diamond enlarged here for visibility.

Experiment 2 Results & Discussion

Participants were significantly faster in the photo preview condition (M=777 ms) compared to the rectangle preview condition (M=814 ms; t(99) = 4.36, p < 0.001, *d* = 0.44; see Figure 5), indicating the presence of photo "scene priming" effects even for a task that does not require any scene layout information or any use of the background scene in the task. Accuracies in the two conditions were high and very similar (rectangle: 98.5%; photo: 98.6%), and a post-hoc uncorrected t-test showed no significant difference between them (t(99) = 0.36, p = 0.70, d = 0.04).

Because square vs. diamond targets are randomly assigned to either target location (with this assignment counterbalanced across participants for each image), the effects here cannot be the result of layout information being predictive of the locations of squares vs. diamonds, or of the visual features of these targets, or of the response participants need to make. Instead, the results support the hypothesis that scene priming with photograph previews can result from participants being faster to localize the probes; in other words, that response times are facilitated by the sudden onsets of the probe shapes when detailed visual information is shared by the preview image and the target image. Because the preview images do contain layout information, we cannot rule out the hypothesis that participants obligatorily process this information. However, because of the absence of a relationship between the layout of the scene and the shape task, there is no plausible explanation for how faster processing of the background scene's layout could speed shape judgments. Further, overall faster reaction times in this experiment compared to Experiment 1 are consistent with the task in the current experiment not requiring any processing of the background scenes. (By contrast, in Experiment 1, once the dots were localized in each condition, a depth task also needed to be performed.)

This hypothesis that the source of scene priming effects may be the detection of the onset of the target shapes provides a potential explanation for the lack of scene priming in the line drawings we used in Experiment 1. That is, while the sparse line drawings contained significant depth information, they were more abstract and considerably less visually detailed than Sanocki & Epstein's (1997) line drawings, causing them to share less low-level visual information with the target images. Thus, it may be that this lack of visual detail prevented participants from detecting the onset of the probes efficiently. To examine this hypothesis, we next sought to test the source of the scene priming effects found using Sanocki & Epstein's (1997) original stimuli, and in particular the robustness of these effects to visual masking and increased delay, both of which should severely curtail participants' ability to quickly detect the onset of the probes if such detection relies on iconic memory (Irwin & Thomas, 2008).

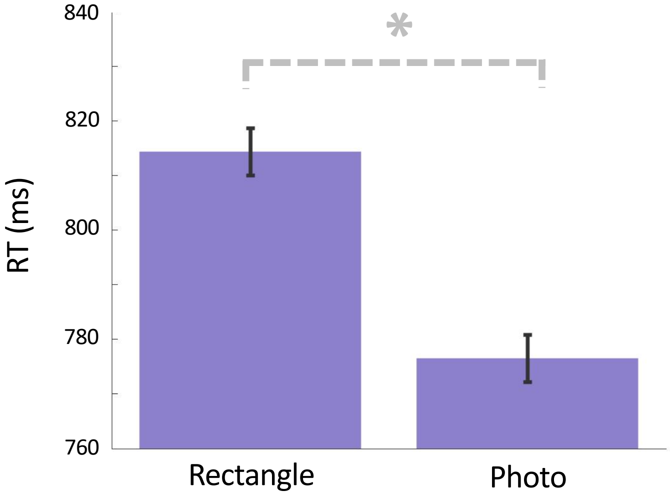


Figure 1.5: Means of reaction times in each preview condition in Experiment 2. Error bars are within-participant SEM. N = 100.

Experiment 3: Replication using original Sanocki & Epstein stimuli; effects abolished using 200ms masked delay period

In Experiment 3, we asked what type of memory store drives scene priming effects. Since these effects may be dependent on the amount of visual detail present shared between preview images and target images, and appear to occur even when the background scene is irrelevant, this raises the possibility that they could arise from integration between the preview and the target scene and the improved ability of participants to detect the probes that results from this integration. Thus we hypothesized that they may be driven not by a robust working memory representation but by a high-capacity but fragile visual memory like iconic memory.

A classical distinction in visual memory is between iconic memory and visual working memory, with high-capacity sensory memory ("iconic" memory) decaying quickly and being easily disrupted by masks, and visual short-term memory being relatively robust to longer delays and visual masks (Irwin & Thomas, 2008). Thus, we reasoned that if the benefits of detailed line drawing previews and photograph previews arose from integration between the preview scene and the target scene in iconic memory, this memory should be interrupted by a visual mask and/or by a longer delay period, even if this delay period remains quite short. By contrast, if the preview benefit reflects a head-start in scene layout processing or participants' ability to hold scene layout in working memory, the preview benefit should remain even after a brief visual mask and a 200ms delay.

Thus, using Sanocki & Epstein's original (1997) stimuli and timing, we first replicated both the photo preview benefit and the line drawing preview benefit. Critically, we included two delay period conditions: an un-masked delay period of the same duration as the original experiments (87 ms) and a masked delay period of 200ms. If Sanocki & Epstein's scene priming effects were driven by information held in iconic memory, the mask and the longer delay between the preview and target image should abolish the preview benefits. On the other hand, if scene priming effects are driven by information in a more robust form of visual memory, such as visual working memory, the scene priming benefits should remain.

Experiment 3 Method

The design, set size and analysis plan for this experiment were preregistered (pre-registration for this experiment here: https://aspredicted.org/rk6f6.pdf; see Supplemental Materials for all pre-registrations).

Participants: Participants were 306 Mechanical Turk workers (102 in each counterbalance condition) who participated in exchange for monetary compensation. We sought (and preregistered) greater power in this experiment as we were predicting a smaller or absent effect of scene previews in the masked conditions.

Stimuli: Stimuli were the original Sanocki & Epstein (1997) target images, scene photographs, and line drawings. The rectangle preview was created in Matlab. In addition to these 3 preview conditions, which we focus on here, the experiment also contained mirrored line drawing previews, as our original interest was to examine the role of spatial layout vs. more global scene properties in scene priming (see also Experiment 1 replications in the Appendix). In this experiment, we do not focus on the mirrored line drawing condition because in this particular set of stimuli the images are extremely symmetrical (with only the exception of the pool image), and thus there is no real difference in the informativeness of the original line drawings and the mirrored line drawings (see Figure A7).

Design & procedure: See Figure 6 for example trial. Preview conditions were blocked, with the order of blocks counterbalanced across participant groups using a balanced latin square. In this experiment, following Sanocki and Epstein (1997), participants task was to judge which of two chairs was closer in depth to the viewer (rather than the red dots in the previous experiments).

Analyses: Analyses and exclusion criteria were the same as for Experiment 1, except that we were now also interested in how any line drawing or photo preview benefits changed according to mask condition. Our preregistered exclusion criteria resulted in the exclusion of 17 participants (15 for accuracy, one of whom also had too many trials faster than 150ms; there were also 2 participants with median RTs slower than 2 seconds in at least one condition).

Experiment 3 Results & Discussion

In the un-masked condition, we found benefits for both line drawings (M=807 ms) and photographs (M=800 ms) over rectangle previews (M=826 ms; line drawings vs. rectangles: t(305) = 3.18, p < 0.002 *d* = 0.18; photos vs. rectangles: t(305) = 3.88, p < 0.001, d = 0.22; see Figure 7). However, both effects were abolished in the masked condition (line drawings vs. rectangles: t(305) = -1.26, p = 0.21, d = -0.07; photos vs. rectangles: t(305) = -0.56, p = 0.57, d = -0.03), with the direction of means for both being in the direction of the preview slowing response, and with Bayes factors showing substantial evidence favoring the null hypothesis in both cases (Scaled JZS Bayes Factor = 7.1 line drawing vs. rectangles; 13.4 photos vs. rectangles; using default of r=0.707 and the method of Rouder, Speckman, Sun, Morey, & Iverson, 2009). A post-hoc power analysis suggests that if the preview benefits in the masked condition were of the same effect size as in the unmasked conditions (~d=0.20), we had 96.7% power to detect this in the current study with our sample size. Comparing the benefit in the masked vs. unmasked conditions, both drawing vs. rectangle and photo vs. rectangle were significantly smaller in the masked compared to the un-masked conditions (line drawing benefit: t(305) = 2.97, p = 0.003, d = 0.17; photo benefit: t(305) = 3.02, p = 0.003, d = 0.17). Mean accuracies in each combination of mask and preview condition ranged between 98.0% and 98.6%. Post-hoc uncorrected t-tests showed no significant differences in any pairs of conditions within either mask/delay condition, or for either of the two critical interactions across mask/delay conditions.

Note that in this experiment using the Sanocki and Epstein (1997) stimuli, rather than making a depth judgment on a pair of red circles, participants had to make a depth judgment on two large chairs that appear in the target scene but are not present in the previews. Thus, the raw reaction times are numerically faster than in Experiment 1, likely reflecting easier localization of the larger chair targets compared to the smaller dot targets. The faster overall reaction times in the 200ms masked condition are consistent with participants benefiting from a longer preparation time compared to the 80ms no-mask condition. While this possibility does not detract from our main conclusions, it prevents us from making any additional conceptual claims based on the overall RT differences in the 80ms no-mask condition vs. the 200ms masked condition. Because the reaction times in our study are well within the range reported for previous scene priming effects (as fast as 562 ms in Sanocki & Epstein, 1997 and as slow as 1029 ms in Sanocki, 2013), this argues that the lack of scene priming in our masked condition is not due to ceiling effects.

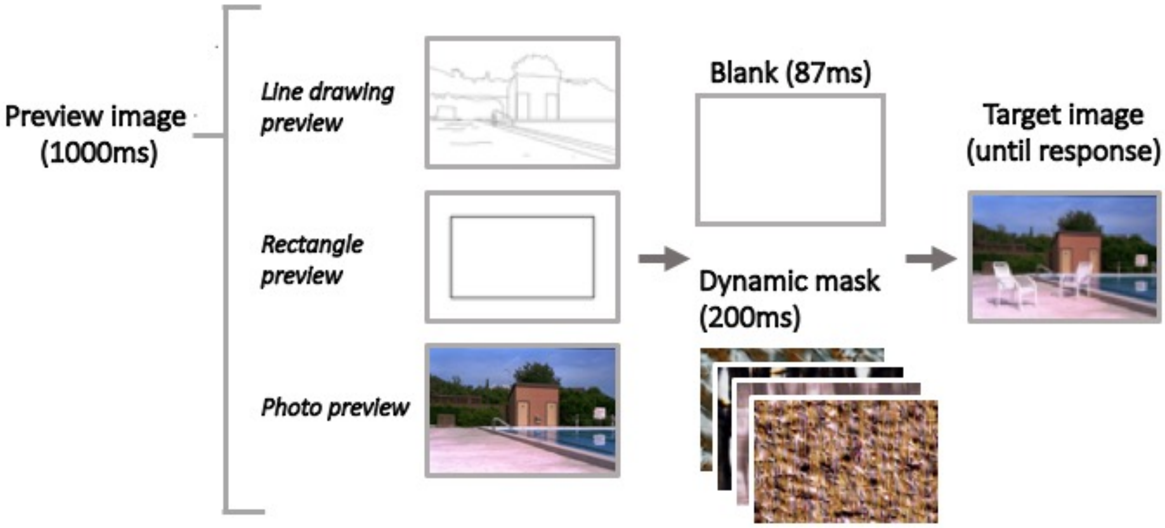


Figure 1.6: Trial timing and conditions for Experiment 3. The line drawing and photo previews do not have the chairs present that are present in each of the target images, and the judgment required on the target image is which of two chairs would be closer to the viewer in depth in real life. In the task, first, a preview image appeared for 1000ms. It was either followed by an 87ms blank, as in the first two experiments (and as in Sanocki & Epstein, 1997), or a dynamic visual mask, for 200ms. Preview and target images were the same as in Sanocki & Epstein (1997).

The fact that both effects were abolished by a longer but still short (200ms) delay and a mask argues that the original preview benefits were due to visual information held in high-capacity sensory memory (e.g., iconic memory). Because a wide variety of information can be stored in iconic memory, including low-level visual information such as patterns of orientation across an image, the results of the present experiment further argue that scene priming paradigms are not able to isolate the effects of scene layout information stored across a delay period. Instead, these results are also consistent with the interpretation that preview images facilitate participants' search for the probes rather than giving them a head-start on layout processing.

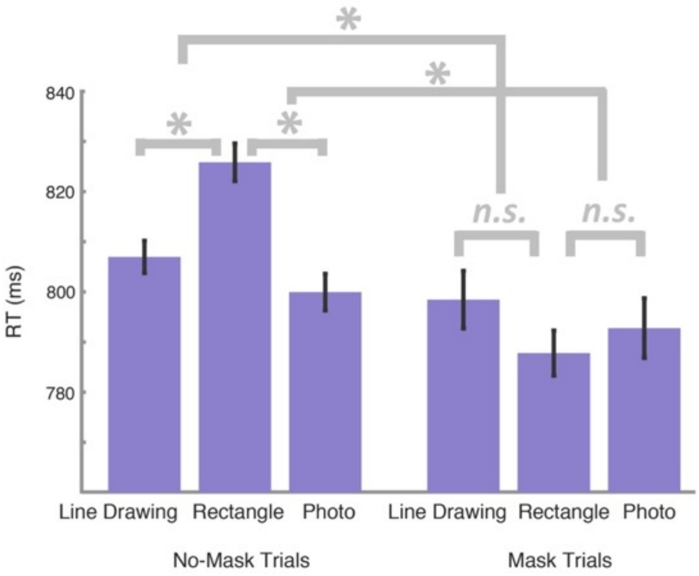


Figure 1.7: Reaction times in each preview condition in Experiment 3. Bars represent means over all participants. Error bars are within-participant SEM. N = 306.

General Discussion

In three experiments, we showed that the effects of scene previews on subsequent depth judgments (termed 'scene priming'; Sanocki & Epstein, 1997) are: (1) present for visually detailed preview images, but not for sparser preview images that still contain depth information; and (2) are driven by information held in iconic memory or another short-term and maskable memory store. In particular, we showed that while both photograph previews (Experiments 1 and 3) and visually detailed line drawings (Experiment 3) produced scene priming benefits, abstract line drawings (containing only the boundaries of major objects and surfaces; Experiment 1) did not, despite containing significant depth information. This is not what we would expect if scene previews facilitated performance by giving participants a head start on layout processing. Further arguing against the idea that scene previews primarily facilitate layout processing, we found a photograph preview benefit even for a task in which the background scene was completely irrelevant (Experiment 2). Finally, we found that the scene priming effects from Sanocki &

Epstein's original (1997) photographs and detailed line drawings both disappeared when the delay period is masked, suggesting that scene priming effects are driven by information held in iconic memory. Together, our data suggest that scene previews may primarily speed participants' localization of the probe shapes on the target image.

Relationship to Previous Scene Priming Findings: Our results are in line with previous studies showing benefits of a scene preview on subsequent processing of a scene. For example, a preview of a real-world scene image facilitates subsequent visual search in that scene (Castelhano & Henderson, 2007; Võ & Henderson, 2010), and both scene photograph and detailed line drawing previews speed subsequent depth judgments on scenes (Sanocki & Epstein, 1997). We consistently replicated photograph preview benefits, and we replicated line drawing preview benefits when using the same line drawings as the original experiment (Sanocki & Epstein, 1997).

However, our results are at odds with the argument that these effects are due to abstract visual information about a scene's layout that speeds participants' judgments by giving them a head start on processing scene layout information. Previous support for this argument is based on based on a few experiments: first, in Sanocki & Epstein (1997), a previewed line drawing of a scene photograph facilitates 3D depth judgments on the photograph. Because the line drawing has less low-level information in common with the target image than a full photograph preview and facilitates depth judgments, they reasoned that layout information is stored across the delay. Second, Sanocki (2003) showed scene priming with moderate retinal shifts between previews and targets (Experiment 5), and Sanocki & Epstein (1997) argue that the viewpoint shifts present in their Experiment 4 are evidence of a more abstract, higher-level representation. Finally, Sanocki (2003; Experiments 2-5) varies lighting direction between preview and target images, disrupting some low-level visual information.

However, while the above experiments show that scene priming benefits persist when some low-level information is varied, the effect is often diminished, and remaining low-level visual information (e.g., the orientation information present in each part of the image) could be driving the preview benefit. Even line drawing previews, which perhaps share the least pixel-by-pixel information with target photographs, still preserve some of the important orientation information in the target photographs, especially the detailed line drawings used in Sanocki & Epstein (1997). Orientation and edge information is well-known to be relevant to scene information. Both local orientations, curvatures and angles (e.g., Walther & Shen, 2014) and the global distribution of orientation information (e.g., Brady et al., 2017; Oliva & Torralba, 2001) are critical to scene recognition. Furthermore, detailed line drawings elicit remarkably similar brain activity in scene regions to real scene photographs (Walther, Chai, Caddigan, Beck, & Fei-Fei, 2011). Thus, it may be that line drawing preview benefits in fact reflect the preservation of these important low-level or mid-level features of a scene that are necessary for participants to notice the onset of a new set of objects, rather than reflecting the representation of more abstract properties such as spatial layout.

Another study using scene previews (Castelhano & Pollatsek, 2010) shows the limited viewpoint tolerance of scene priming effects, and it is notable that the viewpoints that give the largest scene priming benefits are also the ones with the most low-level overlap with the target images. This is in line with the results we report here. Prior work by Gottesman (2011) has the potential to demonstrate the maintenance of more abstract information from scene previews, but the conclusions of that work rest on the particular details of the stimuli they used and how specific the effects of boundary extension are to higher levels of the visual hierarchy. Future work could investigate the potential of their paradigm for specifically investigating scene layout information stored in memory.

Our findings are also consistent with arguments made in Germeys & d'Ydewalle (2001), but while their results call into question scene priming results with significant pixel-by-pixel overlap between preview and target images, ours argue that even studies designed to reflect a more abstract memory store, such as those using line drawings as previews (e.g., see Sanocki 2003), may instead be picking up on the speedier detection of target shapes.

Implications for Representations of Space in Visual Memory: There is a long-running and broad debate over how much information we maintain about the world in memory (O'Regan & Noë, 2001), whether and when we are able to integrate information from successive fixations into a more complete picture of our surroundings (Henderson, 1997; Irwin, Yantis, & Jonides, 1983; Irwin, 1991), and what format these representations are in. Investigating the types of scene information retained in memory has the potential to shed light on how much information we maintain in memory about the world and how we combine information across successive fixations to build a more complete picture of our surroundings. While a good deal of work has been done on the maintenance of object information across brief delays and eye movements, less is known about whether scene layout information persists across eye movements, and if so, how this type of memory fits into the process of maintaining a stable representation of the world. The flash-preview-moving-window paradigm (Castelhano & Henderson, 2007; Võ & Henderson, 2010) demonstrates memory for a size-invariant representation of some information about a natural scene, but it is unclear what the content of this representation is. Change blindness effects (Carlson-Radvansky & Irwin, 1995; Franconeri & Simons, 2003; Luck & Vogel, 1997; Grimes, 1996; McConkie & Currie, 1996; Phillips, 1974; Rensink, O'Regan, & Clark, 1997; Simons, 1996) argue that when we are unable to rely on iconic memory (as is often the case in the real world, visual details are often lost). It is an important question the extent to which people store detailed spatial layout information in memory — and particularly working memory, which is quite capacity limited. Because the current findings call into question one of the main literatures used to support the existence of spatial layout representations, it remains an open question the extent of the layout of specific surfaces in a scene (scene layout) that we are capable of maintaining in working memory.

One of the challenges for future work is understanding how scene layout representations can be quantified and incorporated into existing models of working memory. In particular, while working memory is known to be quite capacity limited, there is significant debate in the visual working memory literature about whether the units of working memory capacity are discrete "slots" or a more continuous resource that can be used to remember fewer objects with more precision or more objects with less precision (Luck & Vogel, 2013; Ma, Husain, & Bays, 2014). Because the layout of a scene is not obviously broken down into discrete objects, it is a challenge to conceptualize how to incorporate it into these primarily object-based models of working memory. Existing models that incorporate both individual objects as well as higher-level information like ensemble structure may be adaptable to incorporate other information like scene layout (Brady & Alvarez, 2011; Brady & Tenenbaum, 2013).

Neural models of working memory more easily accommodate the representation of scene layout information. For example, the occipital place area (OPA) and parahippocampal place area (PPA) are generally seen as perceptual areas, but many neural models of working memory are based on the idea that "perceptual" areas can be recruited for working memory storage (Awh &

Jonides, 2001; Chelazzi, Miller, Duncan, & Desimone, 1993; Curtis & D'Esposito, 2003; D'Esposito, 2007; D'Esposito & Postle, 2015; Harrison & Tong, 2009; Lara & Wallis, 2015; Magnussen, 2000; Miller, Li, & Desimone, 1993; Pasternak & Greenlee, 2005; Serences, Ester, Vogel, & Awh, 2009; Sreenivasan, Curtis, & D'Esposito, 2014). The neuroimaging literature shows evidence of scene-specific representations in perceptual contexts (Dilks, Julian, Paunov, & Kanwisher, 2013; Epstein & Kanwisher, 1998; Maguire, 2001), including boundary information in the OPA (Julian, Ryan, Hamilton, & Epstein, 2016). Thus, future work could examine working memory delay period activity or patterns in these regions to quantify working memory for spatial layout and examine how it interacts with other working memory capacity limits.

Conclusion

The ability to perceive and remember the spatial layout of a scene is critical to understanding the visual world, both for navigation and for other complex tasks that depend upon the structure of the current environment. The present studies offer a new interpretation of scene priming effects, which are one of the primary tools used to study the representation of spatial layout. We find that scene priming effects are driven by visual detail held in iconic memory that does not necessarily isolate scene layout information. Studying scene layout information in memory has the potential to offer fresh insight into several long-standing questions about visual memory, and the current studies are a critical first step towards this goal.

Acknowledgments

The authors thank Ed Vul, Viola Störmer, and John Serences for helpful discussion. This work was supported by an NSF Graduate Research Fellowship to AS and NSF CAREER (BCS-1653457) to TFB. Chapter 1, in full, is a reprint of the material as it appears in: Shafer-Skelton, A., & Brady, T. (2018). Scene layout priming relies primarily on low-level features rather than scene layout. The dissertation author was the primary investigator and author of this paper.

Appendix

Figures A1-A3: Preview benefits by participant

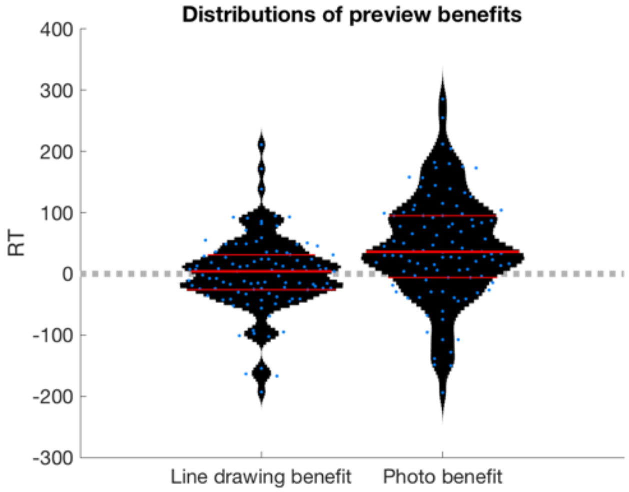


Figure 1.A1: Distributions of individual participants' line drawing and photo preview benefits in Experiment 1. Red lines mark the boundaries of quartiles, and blue points are individual participants' preview benefits in each condition. Note that because we collected many participants but with relatively few trials per participant (to avoid repeating scenes too often), the spread of participants data is larger than in a typical psychophysics study, whereas our power to estimate the grand average across participants and the variation across participants is higher.

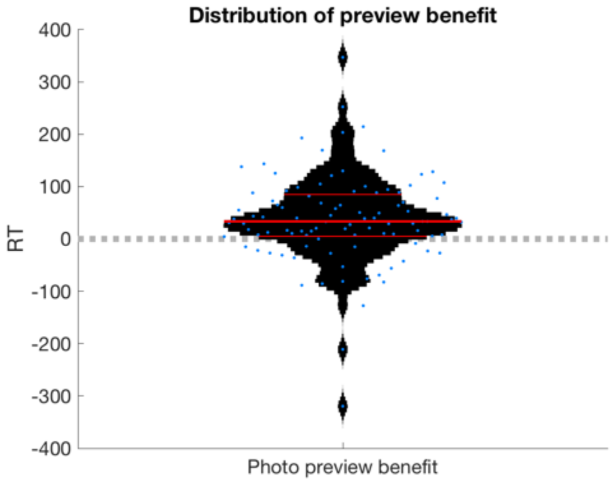


Figure 1.A2: Distributions of individual participants' photo preview benefits in Experiment 2. Red lines mark the boundaries of quartiles, and blue points are individual participants' preview benefits.

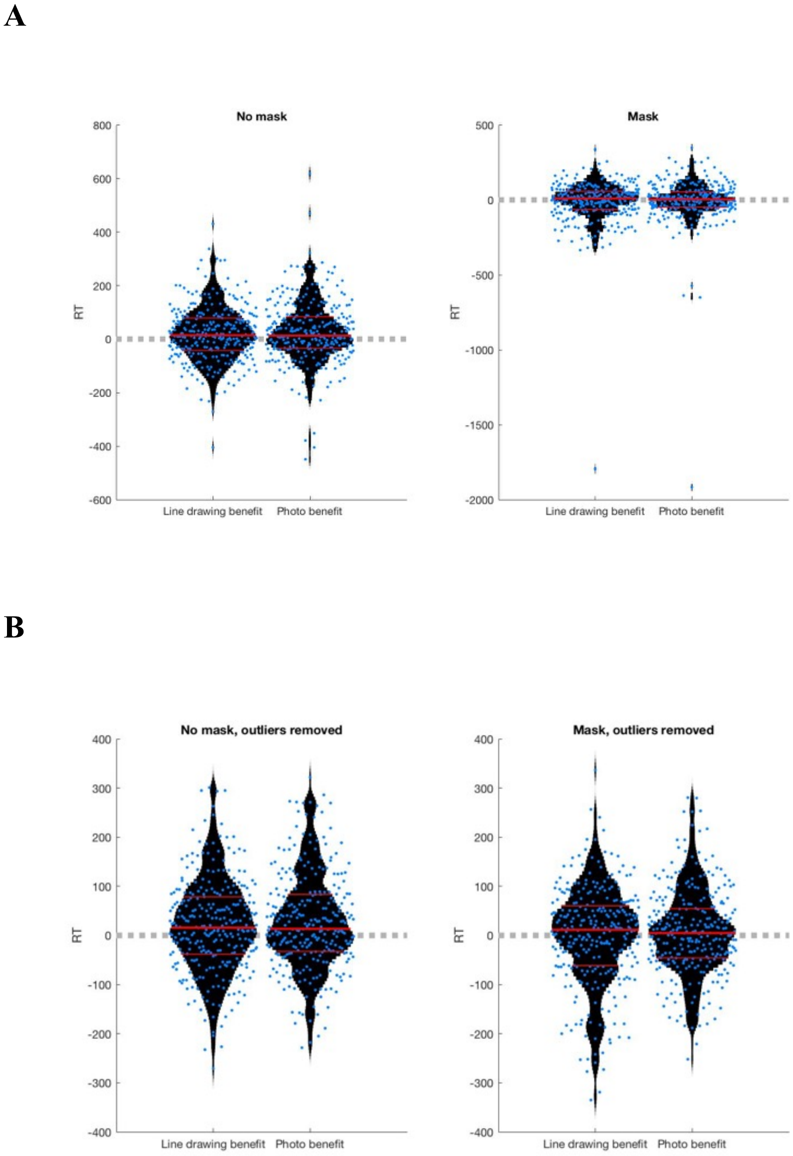


Figure 1.A3: (A) Distributions of individual participants' photo and line drawing preview benefits in Experiment 3 show a few outliers. We identified outliers as any participants who had a median RT in any condition that was three standard deviations more extreme than the mean. Supplemental analyses show that post-hoc removal of these outliers gives the same pattern of results for our main analyses: line-drawing preview benefit, no mask: t(295) = 3.43, p < 0.001 , d = 0.20; photo preview benefit, no mask: t(295) = 4.88, p <0.001 , d = 0.28; line-drawing preview benefit, mask: t(295) = -0.22, p = 0.83 , d = -0.01; photo preview benefit, mask: t(295) = 1.45, p = 0.15, d = 0.08; line-drawing benefit diminishes with mask: t(295) = 2.96, p = 0.003, d = 0.17; photo benefit diminishes with mask: t(295) = 2.80, p = 0.005, d = 0.16. (B) Distributions of preview benefits with outliers removed.

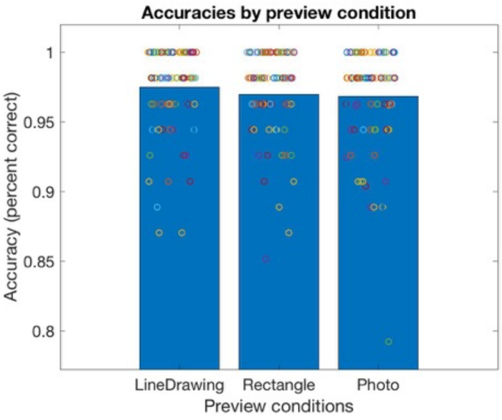


Figure 1.A4: Accuracy data for Experiment 1. Circles are individual participants.

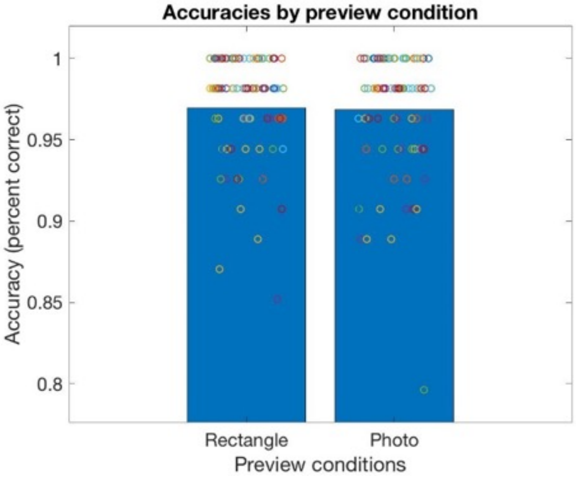


Figure 1.A5: Accuracy data for Experiment 2. Circles are individual participants.

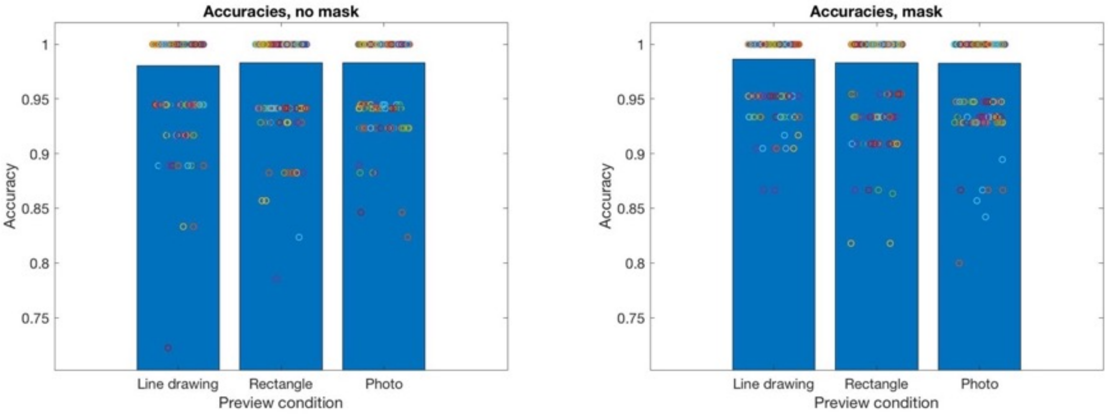


Figure 1.A6: Accuracy data for Experiment 3. Circles are individual participants.

Experiment A1: Verifying sparse line drawings contain layout information

The design, set size, and analysis plan for this experiment were preregistered (see below for pre-registrations).

Participants: Participants were 100 Mechanical Turk workers who participated in exchange for monetary compensation. No participants participated in any other experiments using these line drawings.

Stimuli: Stimuli were the line drawing images used in Experiments 1 and 2, with target dots placed on them in the locations corresponding to the photo target images from Experiments 1 and 2.

Design and procedure: In this experiment, there were no preview images, and participants saw each target line drawing once. During practice, participants were shown examples of line drawings created from photographs, and they practiced choosing which dot would indicate the closer part of the line drawing if the scene existed in three dimensions. Participants were given feedback for correct and incorrect answers in the practice, but only for incorrect answers during the main experiment.

Analyses: In this experiment, we analyzed average performance as well as performed a two-tailed binomial test on each image to determine whether participants' depth judgments were significantly above chance.

Results: Participants were 67% accurate at this task, significantly above chance (t(99) = 17.46, p < 0.001, *d* = 1.75). We found that 35 of the 54 images had above-chance depth judgments in the binomial test, and these are the images that are the focus of the post-hoc analysis in Experiment 1.

Experiment A2: Mirrored and un-mirrored line drawing previews

The design, set size, and analysis plan for this experiment were preregistered (see below for pre-registrations).

Participants: Participants were 100 Mechanical Turk workers who participated in exchange for monetary compensation. No participants participated in any other experiments using these line drawings.

Stimuli: Stimuli were the same as in Experiment 1, except there was an additional preview condition using left/right mirror-reversed line drawings, which were created using Matlab.

Design and procedure: In this experiment, there were four preview conditions: line drawing preview, mirrored line drawing preview, uninformative rectangle preview, and photo preview. The order images appeared in was randomized with the constraint that each target image was presented for the first time before any images were presented for the second time, for each of four presentations of each image (one per preview condition).

Analyses: Our pre-registered comparison was a t-test between the mirrored line-drawing condition and the un-mirrored line drawing condition. Based on Sanocki & Epstein (1997), we also expected at least the un-mirrored line drawing condition to be facilitated relative to the rectangle baseline condition.

Results and Discussion: We found no significant benefit for either of the line drawing preview conditions compared to the uninformative rectangle baseline (un-mirrored significantly slower than baseline: t(99) = -2.93; p = 0.004; d = -0.29; mirrored no difference: t(99) = -0.70, p = 0.49, *d* = -0.07), making any difference between the two line drawing conditions uninterpretable. We did, however, find a photograph preview benefit (t(99) = 7.66, p < 0.001, *d* = .77), suggesting that the lack of line drawing benefit was not due to participants ignoring previews altogether or lack of trying at the task.

Because of a mistake in counterbalancing, the mappings between condition order and target image was not changed across participants as intended (That is, all participants saw a particular target image first in the photograph condition, and another particular target image first in the un-mirrored line drawing condition, etc).

Experiment A3: Mirrored and un-mirrored line drawings, blocked design

The design, set size, and analysis plan for this experiment were preregistered (see below for pre-registrations).

Participants: Participants were 100 Mechanical Turk workers (25 in each counterbalance condition) who participated in exchange for monetary compensation. No participants participated in any other experiments using these line drawings.

Stimuli: Stimuli were the same as in Experiment A2.

Design and procedure: We reasoned that in Experiment A2 the intermixing of un-mirrored and mirrored line drawings may have caused participants to pay less overall attention to both types of line drawing previews. For this reason, we blocked the preview conditions in Experiment A3. Thus, preview conditions were blocked in this experiment, with the order of blocks counterbalanced across participant groups using a balanced latin square. Other aspects of the design were the same as Experiment A2.

Analyses: Again, our pre-registered comparison was a t-test between the mirrored line-drawing condition and the un-mirrored line drawing condition; based on Sanocki & Epstein (1997), we again expected at least the un-mirrored line drawing condition to be facilitated relative to the rectangle baseline condition.

Results and Discussion: We found no significant benefit for either of the line drawing preview conditions compared to the uninformative rectangle baseline (un-mirrored vs. rect: t(99) = -0.48, p = 0.64, *d* = -0.05; mirrored vs. rect: t(99) = -0.31, p = 0.76, *d* = -0.03), making any difference between the two line drawing conditions uninterpretable. Again, we found a photograph preview benefit (t(99) = 3.08, p = 0.003, *d* = 0.31), suggesting that the lack of line drawing benefit was not due to general inattention to preview images. Because the preview types were blocked and introduced at the beginning of each block, the lack of a line drawing benefit was unlikely to be due to participants ignoring all line drawings because mirrored line drawings were unhelpful.

Experiment A4: Un-mirrored photograph previews facilitate depth judgments better than mirrored photograph previews

The design, set size, and analysis plan for this experiment were preregistered (see below for pre-registrations).

Participants: Participants were 102 Mechanical Turk workers (17 in each of 6 counterbalance conditions) who participated in exchange for monetary compensation. No participants participated in any other experiments using these line drawings.

Stimuli: Target images were the same as in Experiment A2 and A3, and preview images were either rectangle previews, photograph previews, or mirror-reversed photograph previews created in Matlab.

Design and procedure: Preview conditions were blocked in this experiment, with every possible order of blocks equally likely across the 6 participant groups. Other aspects of the design were the same as in Experiments A2 and A3.

Analyses: Our pre-registered comparison was a t-test between the mirrored and un-mirrored photograph preview conditions (note there is a small inconsistency in the pre-registration, which says *line drawings* rather than *photographs* in the analysis section, despite the fact that there were no line drawings in this study); we also expected to replicate the un-mirrored photograph preview benefit we found in Experiments A2 and A3.

Results and Discussion: Our critical analysis found that subjects' median RTs were significantly faster in the un-mirrored photo prime condition compared to the mirrored photo prime condition (t(101) = 2.27, p = 0.026, *d* = 0.22). There was again a benefit of the un-mirrored photo prime compared to the rectangle prime (t(101) = 2.44, p = 0.016, *d* = 0.24) but not for the mirrored photo prime compared to the rectangle prime (t(101) = -0.24, p = 0.81, d = -0.02). This argues against scene priming benefits originating solely from memory for *global properties* (Oliva, 2005) of scenes, such as openness or amount of perspective, since the photograph and mirrored photograph previews had identical global properties, but only the un-mirrored photographs facilitated depth judgments.

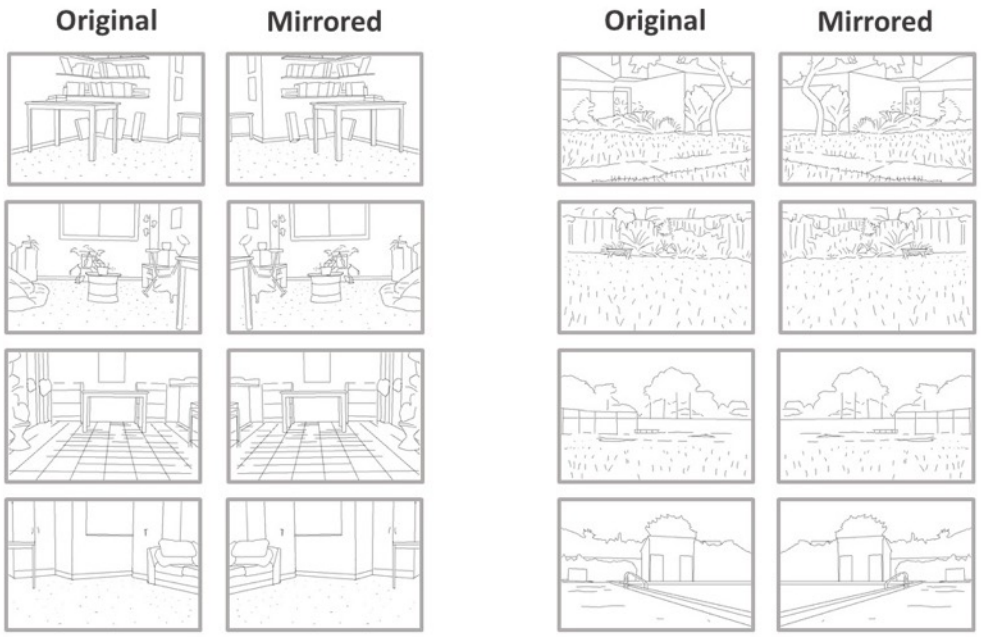


Figure 1.A7: Sanocki & Epstein's original (1997) line drawing stimuli, left columns; mirror-reversed versions of their stimuli, right columns. The images are largely mirror-symmetric, which makes the mirror-reversed line drawing condition in Experiment 3 uninformative.

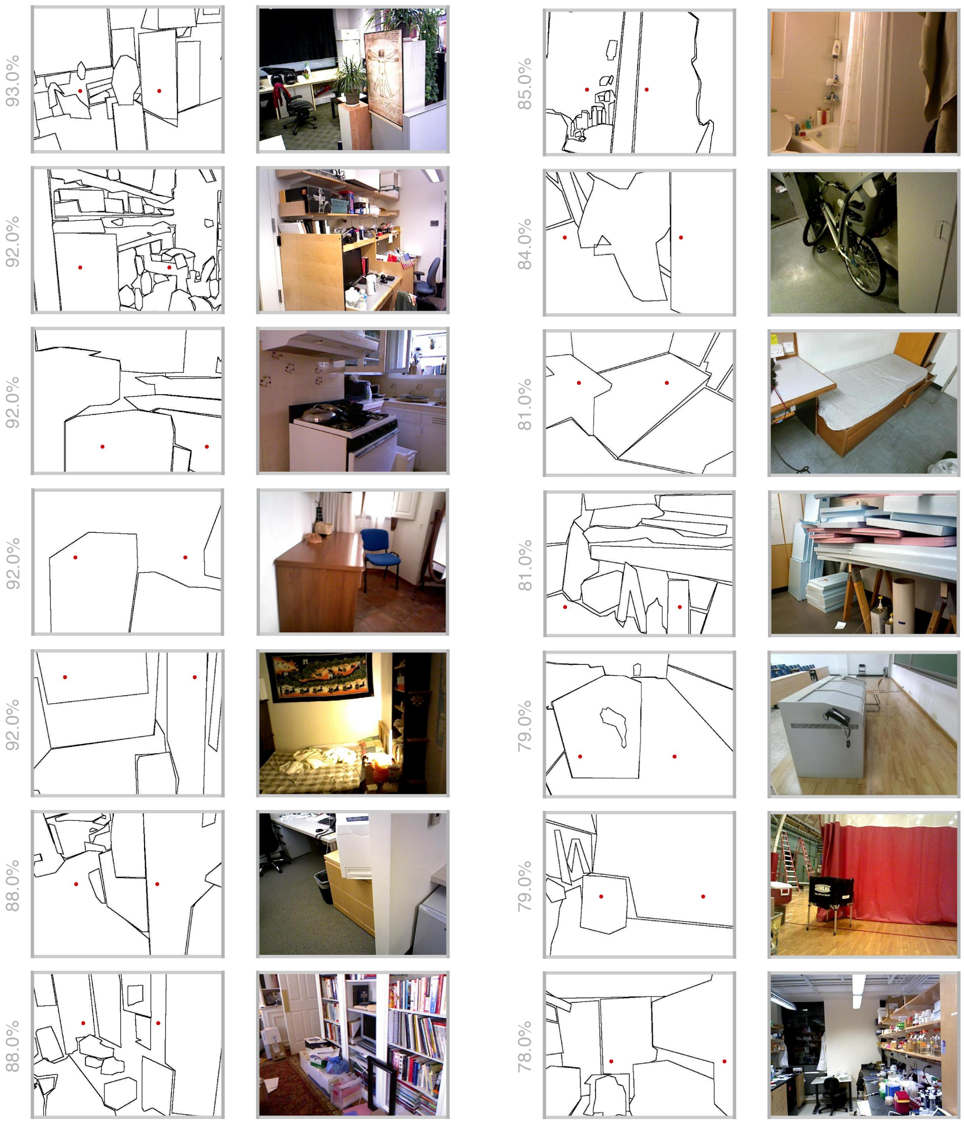


Figure 1.A8: Line drawings and corresponding photographs, Experiment 1 – ordered by depth discrimination accuracy from line drawings alone (indicated next to drawing)

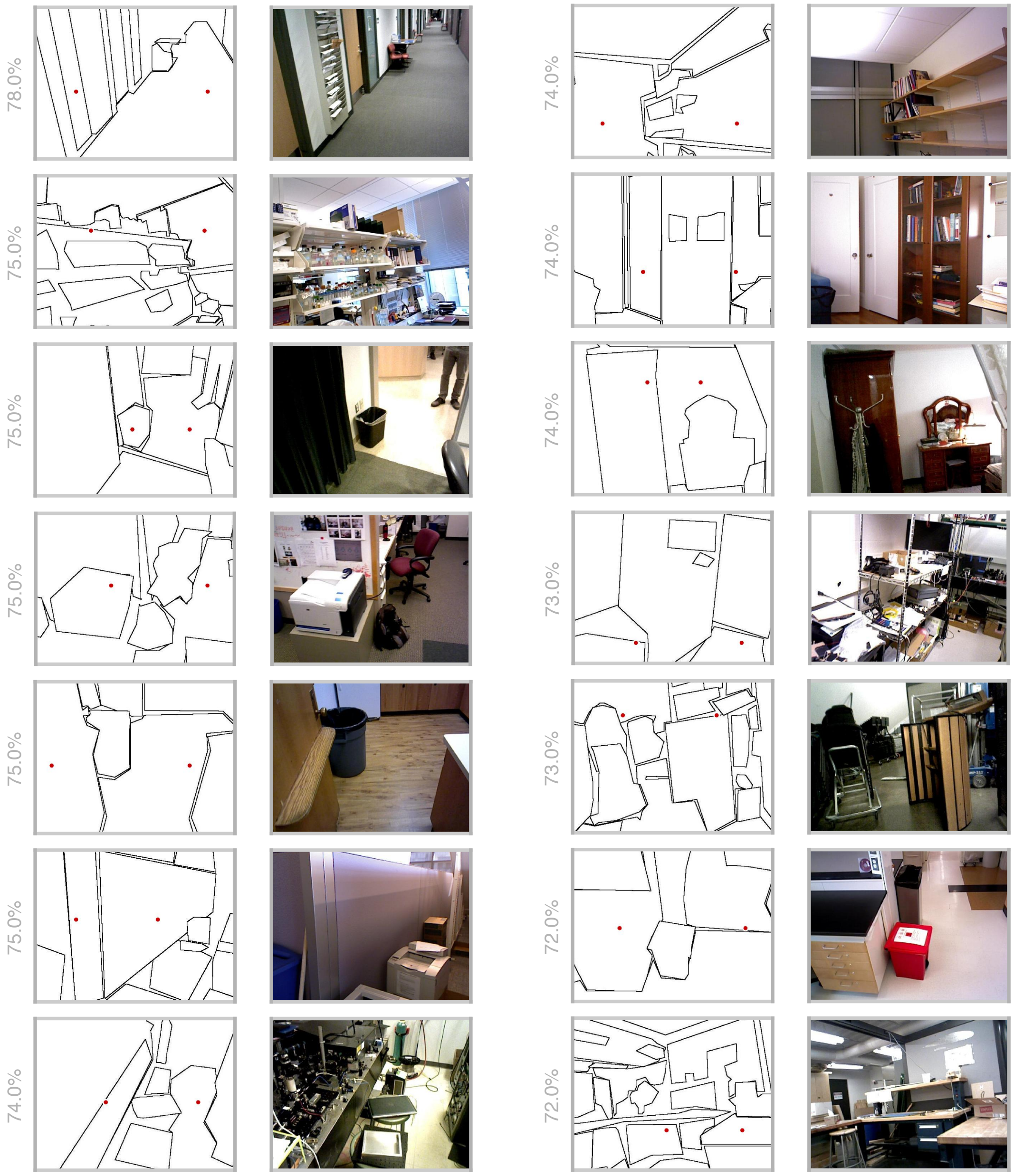


Figure 1.A8: Line drawings and corresponding photographs from Experiment 1 (continued)

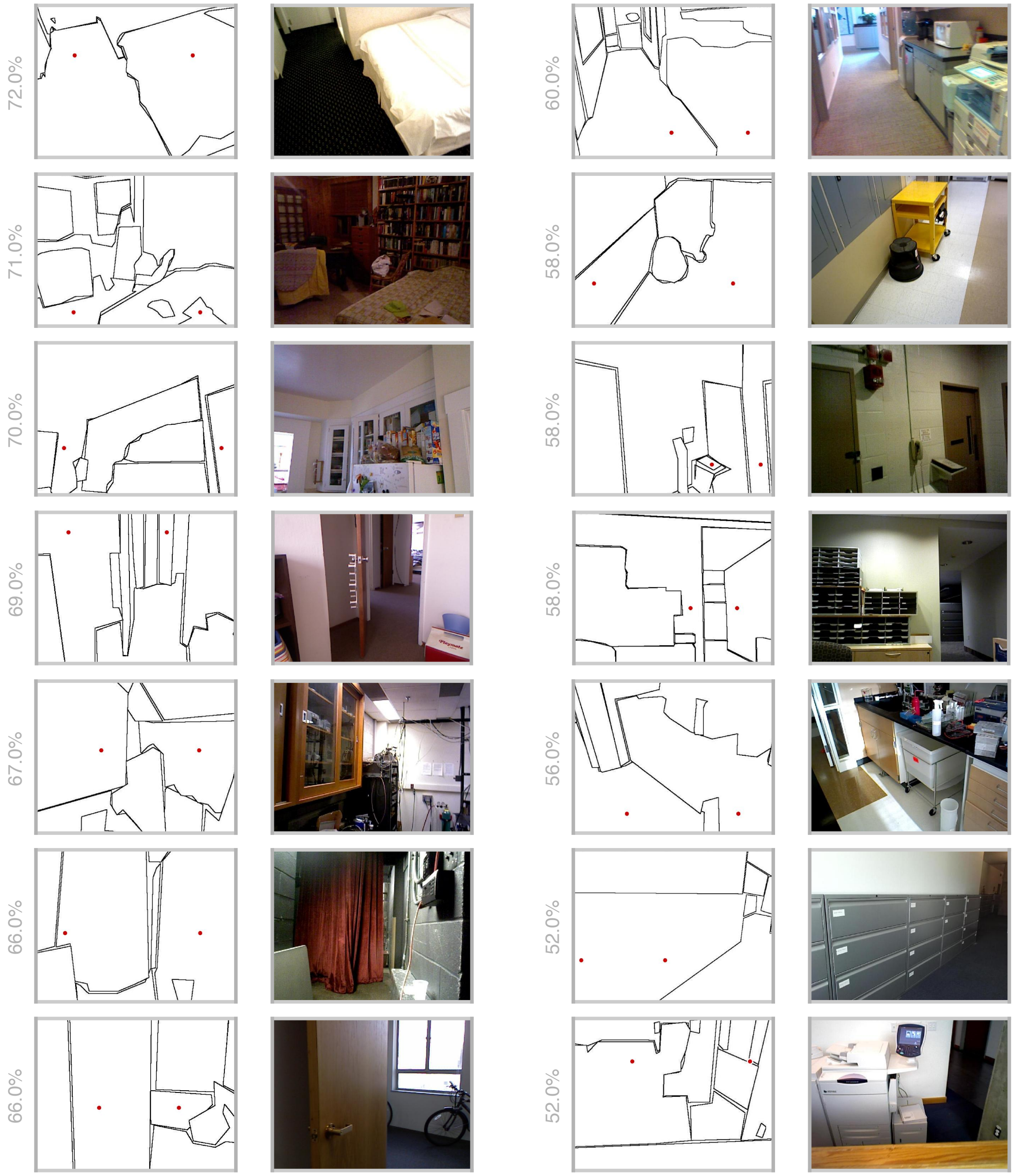


Figure 1.A8: Line drawings and corresponding photographs from Experiment 1 (continued)

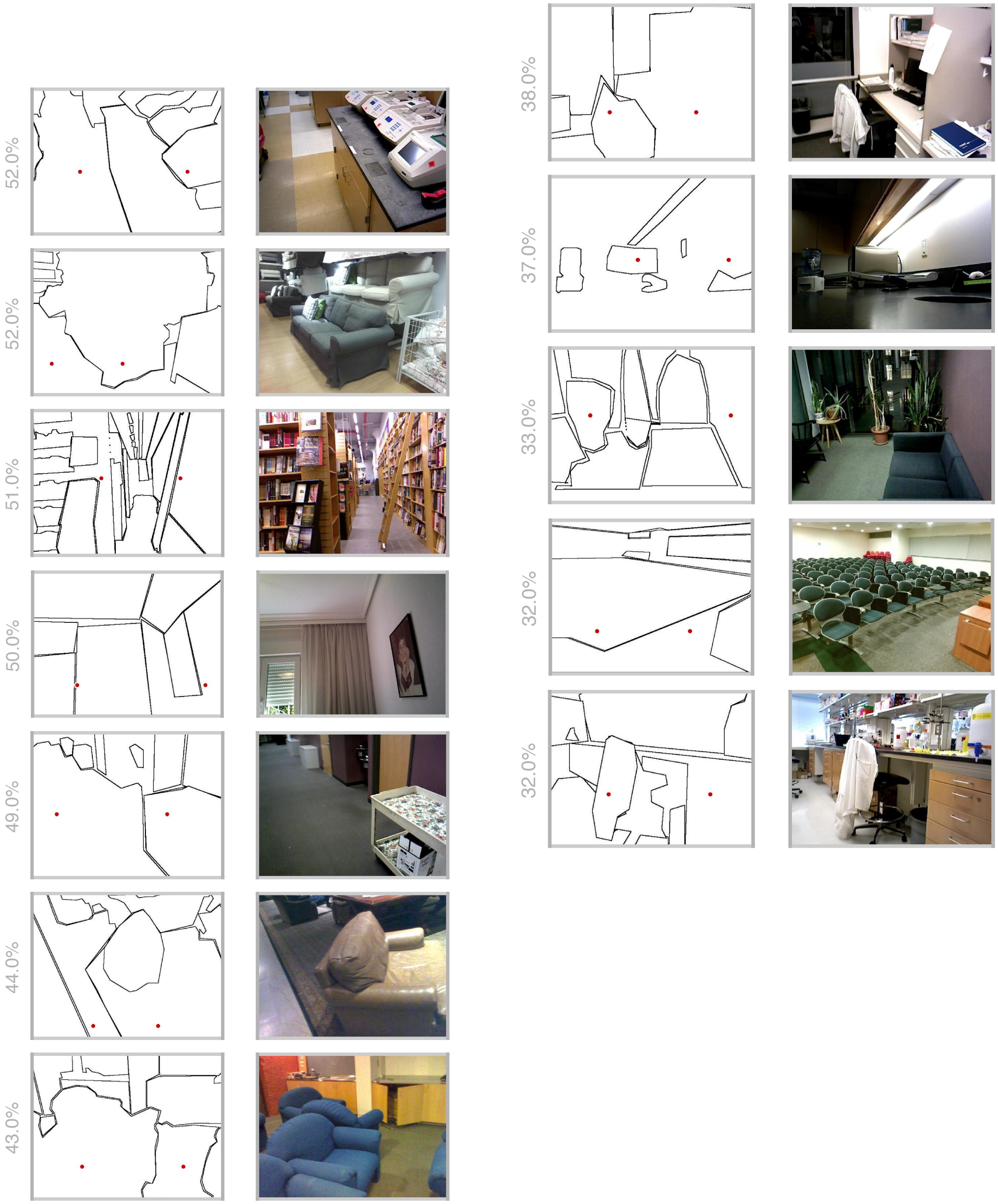


Figure 1.A8: Line drawings and corresponding photographs from Experiment 1 (continued)

References

Awh, E., & Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *Trends in Cognitive Sciences*, *5*(3), 119–126. https://doi.org/10.1016/S1364-6613(00)01593-X

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384–392. https://doi.org/10.1177/0956797610397956

Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 0–17. https://doi.org/10.1037/xhp0000399

Brady, T. F., Störmer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences*, *113*(27), 7459–7464. https://doi.org/10.1073/pnas.1520027113

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, *120*(1), 85–109. https://doi.org/10.1037/a0030779

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk. *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Carlson-Radvansky, L. a, & Irwin, D. E. (1995). Memory for structural information across eye movements. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*(6), 1441–1458.

Castelhano, M. S., & Henderson, J. M. (2007). Initial Scene Representations Facilitate Eye Movement Guidance in Visual Search. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 753–763. https://doi.org/10.1037/0096-1523.33.4.753

Castelhano, M. S., & Pollatsek, A. (2010). Extrapolating spatial layout. *Memory & Cognition*, *38*(8), 1018–1025. https://doi.org/10.3758/MC.38.8.1018

Chelazzi, L., Miller, E. K., Duncan, J., & Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature*, *363*(6427), 345–347. https://doi.org/10.1038/363345a0

Chun, M. M., & Jiang, Y. H. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognit Psychol*, *36*(1), 28–71.

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*. https://doi.org/10.1016/S1364-6613(03)00197-9

D'Esposito, M. (2007). From cognitive to neural models of working memory. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 362, pp. 761–772). https://doi.org/10.1098/rstb.2007.2086

D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual Review of Psychology*, *66*, 115–42. https://doi.org/10.1146/annurev-psych-010814-015031

Di Lollo, V. (1980). Temporal Integration in Visual Memory. *Journal of Experimental Psychology: General*, *109*(1), 75–97. Retrieved from http://wexler.free.fr/library/files/di lollo (1980) temporal integration in visual memory.pdf

Dilks, D. D., Julian, J. B., Paunov, A. M., & Kanwisher, N. (2013). The Occipital Place Area Is Causally and Selectively Involved in Scene Perception. https://doi.org/10.1523/JNEUROSCI.4081-12.2013

Epstein, R. (2005). The cortical basis of visual scene processing. *Visual Cognition*, *12*(6), 954–978. https://doi.org/10.1080/13506280444000607

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601. https://doi.org/10.1038/33402

Eriksen, C. W., & Collins, J. F. (1967). Some temporal characteristics of visual pattern perception. *Journal of Experimental Psychology*, *74*(4 PART 1), 476–484. https://doi.org/10.1037/h0024765

Franconeri, S. L., & Simons, D. J. (2003). Moving and looming stimuli capture attention. *Perception & Psychophysics*, *65*(7), 999–1010. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14674628

Germeys, F., & Ydewalle, G. (2001). Revisiting scene primes for object locations. *Quarterly Journal for Experimental Psychology*, *54A*(3), 683–693. https://doi.org/10.1080/0272498004200036

Gottesman, C. V. (2011). Mental Layout Extrapolations Prime Spatial Processing of Scenes, *37*(2), 382–395. https://doi.org/10.1037/a0021434

Greene, M. R., & Oliva, A. (2009). The briefest of glances: the time course of natural scene understanding. *Psychological Science : A Journal of the American Psychological Society / APS*, *20*, 464–472. https://doi.org/10.1111/j.1467-9280.2009.02316.x

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632–635. https://doi.org/10.1038/nature07832

Henderson, J. M. (1997). Transsaccadic Memory and Integration During Real-World Object Perception. *Psychological Science*, *8*(1), 51–55. https://doi.org/10.1111/j.1467-9280.1997.tb00543.x

Hollingworth, A. (2004). Constructing visual representations of natural scenes: the roles of short- and long-term visual memory. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(3), 519. https://doi.org/10.1037/0096-1523.30.3.519

Hollingworth, A. (2005). The relationship between online visual representation of a scene and long-term scene memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, *31*(3), 396–411. https://doi.org/10.1037/0278-7393.31.3.396

Hollingworth, A., Hyun, J.-S., & Zhang, W. (2005). The role of visual short-term memory in empty cell localization. *Perception & Psychophysics*, *67*(8), 1332–1343. https://doi.org/10.3758/BF03193638

Irwin, D. E., & Thomas, L. E. (2008). Visual Sensory Memory. In *Visual Memory* (pp. 9–43). https://doi.org/10.1093/acprof:oso/9780195305487.003.0002

Irwin, D. E., Yantis, S., & Jonides, J. (1983). Evidence against visual integration across saccadic eye movements. *Perception & Psychophysics*, *34*(1), 49–57. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6634358

Irwin, E. (1991). Information Integration across Saccadic Eye Movements. *Cognitive Psychology*, *23*, 420–456.

Jonides, J., & Yantis, S. (1988). Uniqueness of abrupt visual onset in capturing attention. *Perception & Psychophysics*, *43*(4), 346–354. https://doi.org/10.3758/BF03208805

Julian, J. B., Ryan, J., Hamilton, R. H., & Epstein, R. A. (2016). The Occipital Place Area Is Causally Involved in Representing Environmental Boundaries during Navigation. *Current Biology*, *26*(8), 1104–1109. https://doi.org/10.1016/j.cub.2016.02.066

Kravitz, D. J., Saleem, K. S., Baker, C. I., & Mishkin, M. (2011). A new neural framework for visuospatial processing. *Nature Reviews Neuroscience*, *12*(4), 217–230. https://doi.org/10.1038/nrn3008

Lara, A. H., & Wallis, J. D. (2015). The Role of Prefrontal Cortex in Working Memory: A Mini Review. *Frontiers in Systems Neuroscience*, *9*. https://doi.org/10.3389/fnsys.2015.00173

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2013.06.006

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356. https://doi.org/10.1038/nn.3655

Magnussen, S. (2000). Low-level memory processes in vision. *Trends in Neurosciences*, *23*(6), 247–251. https://doi.org/10.1016/S0166-2236(00)01569-1

Maguire, E. A. (2001). The retrosplenial contribution to human navigation: A review of lesion and neuroimaging findings. *Scandinavian Journal of Psychology*, *42*, 225–238.

McConkie, G. W., & Currie, C. B. (1996). Visual stability across saccades while viewing complex pictures. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(3), 563–581. https://doi.org/10.1037/0096-1523.22.3.563

Miller, E. K., Li, L., & Desimone, R. (1993). Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *The Journal of NeuroscienceJournal of Neuroscience*, *13*(4), 1460–78. https://doi.org/10.1016/j.conb.2004.03.013

O'Regan, J. K., & Noë, a. (2001). A sensorimotor account of vision and visual consciousness. *The Behavioral and Brain Sciences*, *24*(5), 939-973; discussion 973-1031. https://doi.org/10.1017/S0140525X01000115

Oliva, A. (2005). *Gist of the Scene*. Retrieved from https://s3.amazonaws.com/academia.edu.documents/30821187/oliva04.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1533077700&Signature=uo7Thqhg6Pxgx8OSajOnw%2FpOFH0%3D&response-content-disposition=inline%3B filename%3DGist_of_the_scene.pdf

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene : A Holistic Representation of the Spatial Envelope ∗, *42*(3), 145–175.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*.

Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *The Journal of Neuroscience*, *31*(4), 1333–40. https://doi.org/10.1523/JNEUROSCI.3885-10.2011

Pasternak, T., & Greenlee, M. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience*, *6*, 97–107. https://doi.org/10.1038/nrn1637

Phillips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. *Perception and Psychophysics*, *16*(2), 283–290.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, *8*(5), 368–373. https://doi.org/10.1111/j.1467-9280.1997.tb00427.x

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

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

Sanocki, T. (2013). Facilitatory priming of scene layout depends on experience with the scene. *Psychonomic Bulletin & Review*, *20*(2), 274–281. https://doi.org/10.3758/s13423-012-0332-9

Sanocki, T., & Epstein, W. (1997). Priming Spatial Layout of Scenes. *Psychological Science*, *8*(5), 374–378.

Sanocki, T., Michelet, K., Sellers, E., & Reynolds, J. (2006). Representations of scene layout can consist of independent , functional pieces, *68*(3), 415–427.

Schyns, P., & Oliva, A. (1994). From blobs to boundary edges: evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*. Retrieved from http://pss.sagepub.com/content/5/4/195.short

Serences, J. T. (2016). Neural mechanisms of information storage in visual short-term memory. *Vision Research*, *128*, 53–67. https://doi.org/10.1016/j.visres.2016.09.010

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science*, *20*(2), 207–214. https://doi.org/10.1111/j.1467-9280.2009.02276.x

Simons, D. J. (1996). In Sight, Out of Mind: When Object Representations Fail. *Psychological Science*, *7*(5), 301–305. https://doi.org/10.1111/j.1467-9280.1996.tb00378.x

Song, S., Lichtenberg, S. P., & Xiao, J. (2015). SUN RGB-D: A RGB-D scene understanding benchmark suite. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07–12–June*, 567–576. https://doi.org/10.1109/CVPR.2015.7298655

Sreenivasan, K. K., Curtis, C. E., & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, *18*(2), 82–89. https://doi.org/10.1016/j.tics.2013.12.001.Revisiting

Theeuwes, J. (1991). Exogenous and endogenous control of attention: The effect of visual onsets and offsets. *Perception & Psychophysics*, *49*(1), 83–90. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.457.2361&rep=rep1&type=pdf

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766–786. https://doi.org/10.1037/0033-295X.113.4.766

Võ, M. L.-H., & Henderson, J. M. (2010). The time course of initial scene processing for eye movement guidance in natural scene search. *Journal of Vision*, *10*(3), 1–13. https://doi.org/10.1167/10.3.14

Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., & Fei-Fei, L. (2011). Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proceedings of the National Academy of Sciences*, *108*(23), 9661–9666. https://doi.org/10.1073/pnas.1015666108

Walther, D. B., & Shen, D. (2014). Nonaccidental Properties Underlie Human Categorization of Complex Natural Scenes. *Psychological Science*, *25*(4), 851–860. https://doi.org/10.1177/0956797613512662

Wolfe, J. M., Võ, M. L. H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2010.12.001

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235. https://doi.org/10.1038/nature06860

CHAPTER 2: Deep-net-derived surface estimations of natural scenes predict voxel responses in scene-selective cortex

Abstract

Our world is full of diverse types of visual information, yet visual attention and memory experiments focus almost exclusively on understanding attention/memory of basic visual features and discrete objects, in part because our understanding of how people represent scene surface information cognitively and neurally is very limited. Recent work has made headway in quantifying such surface representations, finding 3D surface information in scene-selective cortex, yet their results were limited to artificially generated images for which ground-truth 3D information exists, and it is unclear if their results would extend to natural scene photographs, which have more naturalistic textures and relationships among visual properties. Here, we use DNNs to estimate ground-truth distance and surface-direction information based only on RGB stimulus images in the publicly available BOLD5000 fMRI data set. Using voxelwise encoding models based on these features, we find that such models can predict significant amounts of activity in scene-selective cortex, suggesting the presence of 3d scene representations when viewing natural images. These results lay the foundation for investigating scene surface processing in more naturalistic environments and tasks, a critical step towards understanding visual and cognitive processes in the real world.

Introduction

As we navigate the world, our visual systems are able to make sense of diverse types of visual information: low- to mid-level visual features like edges or abstract shapes; bodies, facial identities and expressions; letters and words; objects that we may interact with or navigate around; and the large extended surfaces that make up the 3D shape of a scene. One core division in our visual representations is a distinction between the representation of scene surfaces and visual objects. In particular, developmental and comparative work shows differential priority given to representations of major scene surfaces vs. other features across ages and species, with strong dissociations between tasks where children and animals use scene features vs. object features (K. Cheng, 1986; Hermer & Spelke, 1994; Landau & Lakusta, 2009). Furthermore, the observation of anatomically separate regions specialized for processing scenes vs. objects suggests a fundamental divide between representations of simple features or discrete visual objects and visual scenes (e.g., Epstein & Kanwisher, 1998; Park, Brady, Greene, & Oliva, 2011; Silson, Steel, Kidder, Gilmore, & Baker, 2019). These regions can be separately disrupted with TMS (Dilks, Julian, Paunov, & Kanwisher, 2013; Julian, Ryan, Hamilton, & Epstein, 2016; Mullin & Steeves, 2011, 2013), suggesting they play a causal role in processing stimuli of their preferred kind.

Despite the evidence for these differences, considerably more is known about simple visual features and even complex object representations than about scene representations. Importantly, what we know about how attention and memory act on discrete visual objects may not simply extend to the realm of scene representations. For example, several characteristics of scene representations mean they could play different roles than object representations in core cognitive processes; for example, some information about natural scenes is thought to be processed extremely quickly (Greene & Hansen, 2020; Henriksson, Mur, & Kriegeskorte, 2019; Martin Cichy, Khosla, Pantazis, & Oliva, 2017) – in some cases more quickly than, and without recognition of, the objects inside the scene (Greene & Oliva, 2009; Oliva & Torralba, 2006). More distributed properties of scenes can also be processed with reduced focal attention (Alvarez & Oliva, 2009; Groen, Ghebreab, Lamme, & Scholte, 2016), leading to the suggestion that some properties of scenes may be processed via a separate pathway that does not require selective attention, unlike most object processing (Wolfe, Vo, Evans, Greene, 2011).

The asymmetry in what we know about objects and scenes may partly be because more complex stimuli are inherently harder to study (Brady et al., 2019)—naturalistic stimuli with more types of information present leave more alternative explanations to be ruled out. Rather than attempting to perfectly control for representations of lower-level features, recent work has made progress by quantifying variance explained by different feature sets (Greene, Baldassano, Esteva, Beck, & Fei-Fei, 2016; Greene & Hansen, 2020; Groen et al., 2018; Lescroart & Gallant, 2019). This approach has the potential to uncover important sources of variability and lay critical groundwork for more targeted investigations, even though it does not, on its own, perfectly pinpoint the unique contributions of higher-level scene representations (compared to low- and mid-level features).

One important type of scene information that is not well understood is the fine-grained spatial information we use to move through and interact with the world – e.g., the exact location and orientation of major surfaces in scenes. One reason for this may be that it is much more difficult to get ground-truth information for the layouts of major surfaces in scenes than for coarser-grained properties, like whether a scene is indoor or outdoor. A recent study aimed to address this by building virtual 3d environments to use as stimuli (Lescroart & Gallant, 2019;

also see Henriksson, Mur, & Kriegeskorte, 2019), so that the ground-truth 3d information is directly available in the stimuli themselves. They found evidence that scene regions maintained information about scene surfaces and their directions in 3D space, above and beyond local image information like 2D orientation and spatial frequency, making this information a promising potential marker of scene-specific representations.

While this approach offers a substantial degree of flexibility and precise measurement of the underlying ground truth about scene surfaces, the tradeoff is naturalism, with some important real-world characteristics lost—for example, texture and surfaces are manipulated totally independently in these artificial scenes in a way that may not preserve some 2D cues that scene-areas have available to them in the natural world (Lescroart & Gallant, 2019). Thus, in the current work we sought to extend the methods used to isolate 3d scene information by Lescroart and Gallant (2019) to naturalistic scene photographs, rather than 3d renderings. This allows us to test whether such 3d scene information is represented in visual scene regions even in natural stimuli. Another motive for the present work is that the difficulty of creating precise 3d spatial annotations for natural scenes is an unnecessary barrier to studying these navigationally relevant spatial representations, so we also sought to demonstrate an approach that allows us to advance our knowledge of these more complex types of information more quickly and easily than has been possible before, taking advantage of a world where data and stimulus sharing is becoming more common.

To simultaneously address both of these challenges, our approach in the present work was to use pre-trained Taskonomy neural networks (Zamir et al., 2019) to generate 3d features similar to those in Lescroart & Gallant (2019), but using ordinary scene images as a starting point. This provided us with the information needed to fit voxelwise encoding models based on

3d surface features, with prediction success serving as our measure of scene surface "information" present in voxels across cortex.

Methods

fMRI data acquisition and pre-processing: We analyze data from the BOLD5000 data set (Chang et al., 2019), a publicly available fMRI data set. Data were acquired on a 3T Siemens Verio MR scanner at Carnegie Mellon University, using a 32-channel head coil. We use the provided pre-processed data for our analyses, and we analyze the average of the 3rd and 4th TRs after stimulus onset, as those timepoints included the peak responses across all ROIs (Chang et al., 2019).

Participants: We include the three BOLD5000 participants who completed all 15 functional sessions of the experiment, ensuring enough trials for our analyses. This included: CSI1 – male, age 27, right-handed; CSI2 – female, age 26, right-handed; and CSI3 – female, age 24, right-handed.

Task and stimuli: Participants viewed stimuli while fixating, in a slow event-related design, and indicated via button-press how much they liked or disliked each image. From the original 5000+ BOLD5000 stimulus images, we selected the 1000 "Scenes" images (which had been intermixed with other images in each run) that depicted naturalistic scenes, both indoor and outdoor. While these images did not come with any annotations beyond category label, they are chosen to correspond to categories included in the SUN database, an image set used for scene categorization tasks. This ensured a broad sampling of categories (250 total in the BOLD5000 stimuli). There were four exemplars of each category. A small number of images were repeated 3-4 times, and for these, we analyzed only the first presentation.

Regions of interest: We use the provided BOLD5000 region of interest definitions. Eight functional localizer runs were collected over the course of the 15 functional sessions, run at the end of a given session. Stimuli consisted of scenes, objects, and scrambled images, presented in a blocked design, and participants detected repetitions of identical images (one-back task). Scene areas were defined using a scenes > objects+scrambled contrast, and an early visual cortex (EVC) ROI was defined using a scrambled > baseline contrast. The authors note that this resulted in an ROI that may have extended past V1 or V2.

Feature sets: The baseline 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. This is fewer features than in Lescroart and Gallant (2019). To see if this was critical, we also compared fitting success in one subject using a 1,425-feature gabor model similar to Lescroart & Gallant (2019) and found it gave very similar results to the 300-feature model. These features were generated using code from https://github.com/gallantlab/motion_energy_matlab.

To generate the global and quadrant-based 3d scene surface features, we first used Taskonomy pre-trained neural networks (Zamir et al., 2019) to generate the distance to and the surface normals (surface directions in 3d space) of each pixel in each 2d image. We used pytorch to resize each image to 256x256 and used the visualpriors package (Sax et al., 2018) to extract surface-normal and euclidean-distance maps. At this point, we tagged stimuli for exclusion if either corresponding map contained any obvious artifacts (e.g., see Figure 2.1C). This left 978 artifact-free stimulus images. To generate features from these remaining image maps, we adapted code from Lescroart & Gallant (2019). We first chose three distance bins, spaced so that distance bins were roughly equally represented across images (vs. 10 in Lescroart & Gallant, 2019); we included the same 9 surface-direction bins contained in Lescroart & Gallant (2019). In the 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 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 leave-one-session-out cross-validation, resulting in 15 main cross-validation folds. To choose ridge parameters, we nested 10 inner cross-validation folds within the fitting 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 the model for the current main cross-validation fold. There were 13 possible ridge parameters: 0, as well as 12 values logarithmically spaced between 10^{-2} and 10^{5}. Early analyses with this ridge parameter selection matched performance using a more extensive set of ridge parameter choices.

Our main analyses report prediction performance, first averaged across 15 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. For each permutation, we completed the same process of averaging prediction performance across cross-

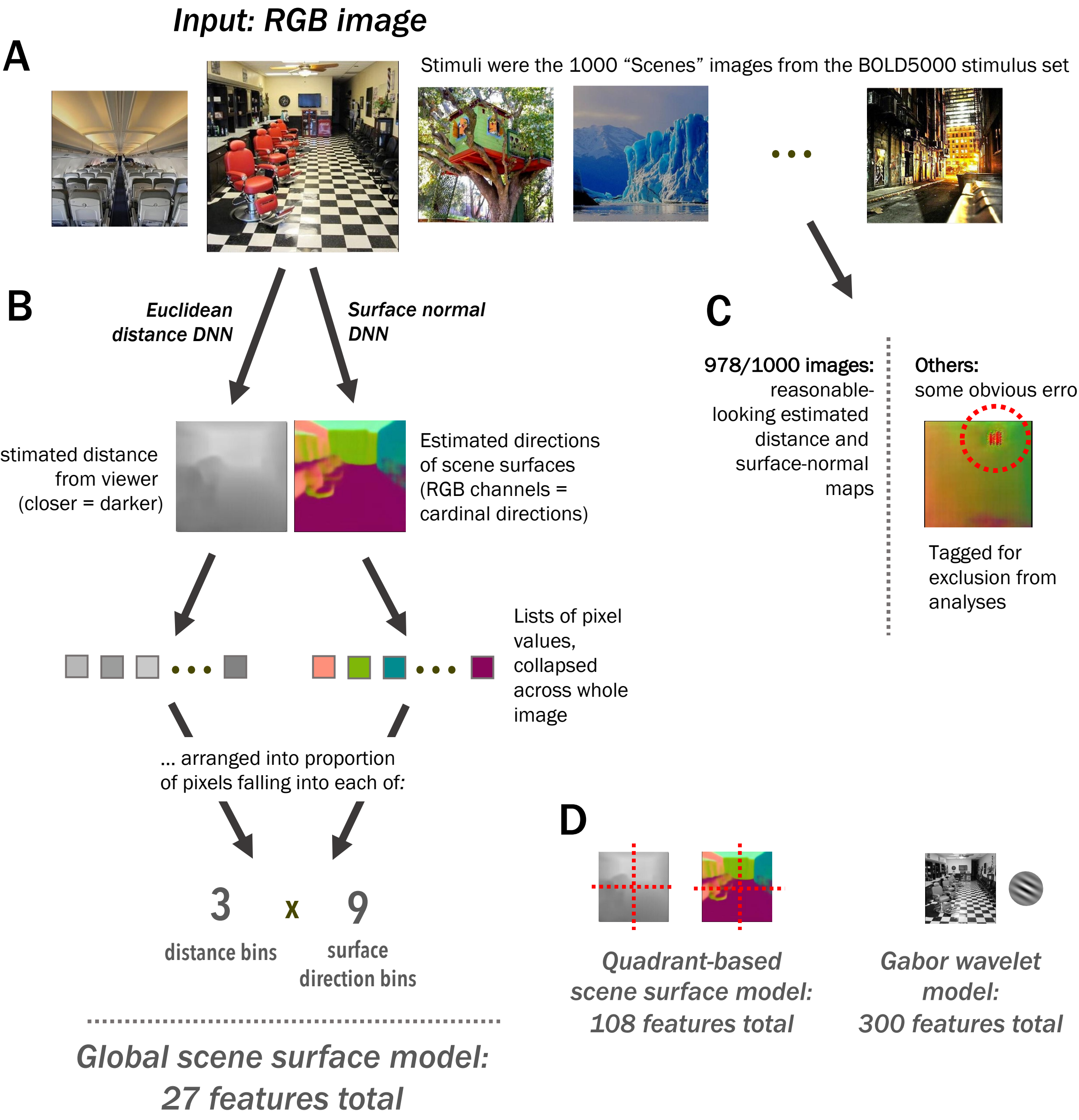

Input: RGB image
A
Stimuli were the 1000 "Scenes" images from the BOLD5000 stimulus set
B
Euclidean distance DNN
Surface normal DNN
Estimated distance from viewer (closer = darker)
Estimated directions of scene surfaces (RGB channels = cardinal directions)
Lists of pixel values, collapsed across whole image
... arranged into proportion of pixels falling into each of:
3
x
9
distance bins
surface direction bins
Global scene surface model: 27 features total
C
978/1000 images: reasonable-looking estimated distance and surface-normal maps
Others: some obvious errors
Tagged for exclusion from analyses
D
Quadrant-based scene surface model: 108 features total
Gabor wavelet model: 300 features total

Figure 2.1. Stimuli and scene surface model features. (A) Stimuli were the 1000 "Scenes" images from the BOLD5000 data set that depicted indoor and outdoor naturalistic scenes (Chang et al., 2019). (B) We used two Taskonomy (Zamir et al., 2019) DNNs to generate estimated distance-from-viewer and surface normal spatial maps. We binned distances into 3 bins and surface normals into 9. Each of the 27 features was the proportion of voxels in an image falling into a particular combination of distance and surface direction bin (Lescroart & Gallant, 2019). (C) Of the 1000 images, the 978 without obvious errors were included in analyses. (D) Left: The quadrant-based scene surface model included 108 features: the 27 combinations of distance and surface-direction bins, for each of the 4 quadrants of images. Right: We used a gabor wavelet model with 300 features to capture spatial-frequency and orientation information.

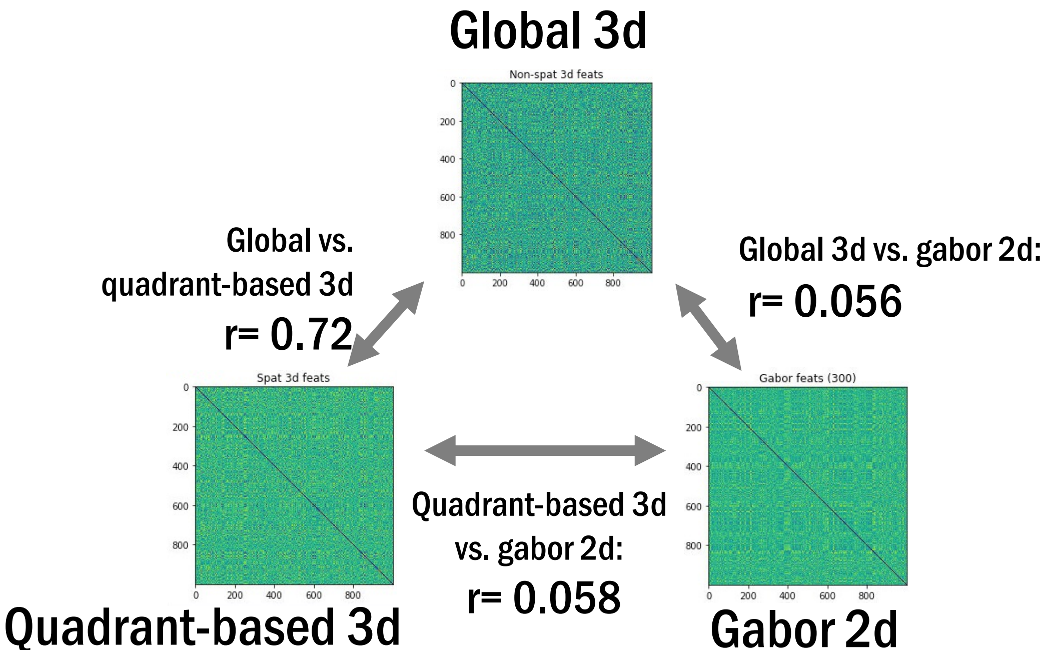


Figure 2.2. RDM correlations for the three feature sets. To create each RDM, artifact images were removed, features were z-scored, and the dissimilarity in each cell was calculated as the cosine distance between the features of the corresponding pair of images. To assess similarity across RDMs, we took the correlation of the cells corresponding unique image pairs. Correlations were very similar across 2 other distance metrics as well.

validation folds and across voxels in a given ROI, generating a null distribution of averaged r-values corresponding to each ROI and participant. We use a significance cutoff corresponding to a two-tailed p-value of 0.05, without assuming a symmetrical null distribution.

Results

Prediction performance for individual models: Consistent with other work, we found the highest prediction performance for the gabor wavelet model in early visual areas (Figure 2.3C; individual subjects' two-tailed p-values via permutation tests: all ps < 0.0001). Performance was lower but also significant in PPA (all ps < 0.0013) and RSC (all ps < 0.019), and significant in 2/3 subjects in OPA (ps < 0.0001, 0.0001, and 0.074). Critically, we were interested in whether a 3d surface-based model, using features estimated with DNNs, could successfully predict performance in scene-selective cortex. Surprisingly, we found significant prediction performance in all three participants and scene-selective ROIs (OPA: all ps < 0.0076; PPA and RSC: all ps < 0.0001). For 2/3 participants, there was also significant prediction performance in EVC, although performance was numerically lower than in scene areas.

We also wondered how the performance of the quadrant-based scene surface model compared to that of the global scene surface model. First, to gauge our potential for success, we wanted to verify that the feature sets were un-correlated enough in this stimulus set—for example, because the BOLD5000 scene stimuli were drawn from the internet, it's possible that images were disproportionately symmetrical compared to scenes we encounter in everyday life. If this were the case, images with similar surface distance/direction features across the whole image (global 3D model) would also have relatively similar features within each quadrant of the image. Even if voxels responded to either global or spatially specific features more strongly, relationships of these features across images would mean that they vary together across trials, limiting our ability to differentiate which features are more responsible for these voxel responses. To assess this, we created representational dissimilarity matrices (RDMs) for both the global and the quadrant-based scene surface model, using cosine distance as our measure of dissimilarity. These RDMs had correlation of 0.72, corresponding to ~52% shared variance across model features in our stimulus set (Figure 2.2). Thus, it could be possible for us to find differences between model performance if they were large enough, but there was enough shared variance in this stimulus set to reduce our potential to find differences. For this data set, we indeed found that quadrant-based model performance (Figure 2.3B) was significant in all scene-selective areas (all ps < 0.0002), with 2/3 participants' EVC areas significantly better than chance (ps < 0.0001,

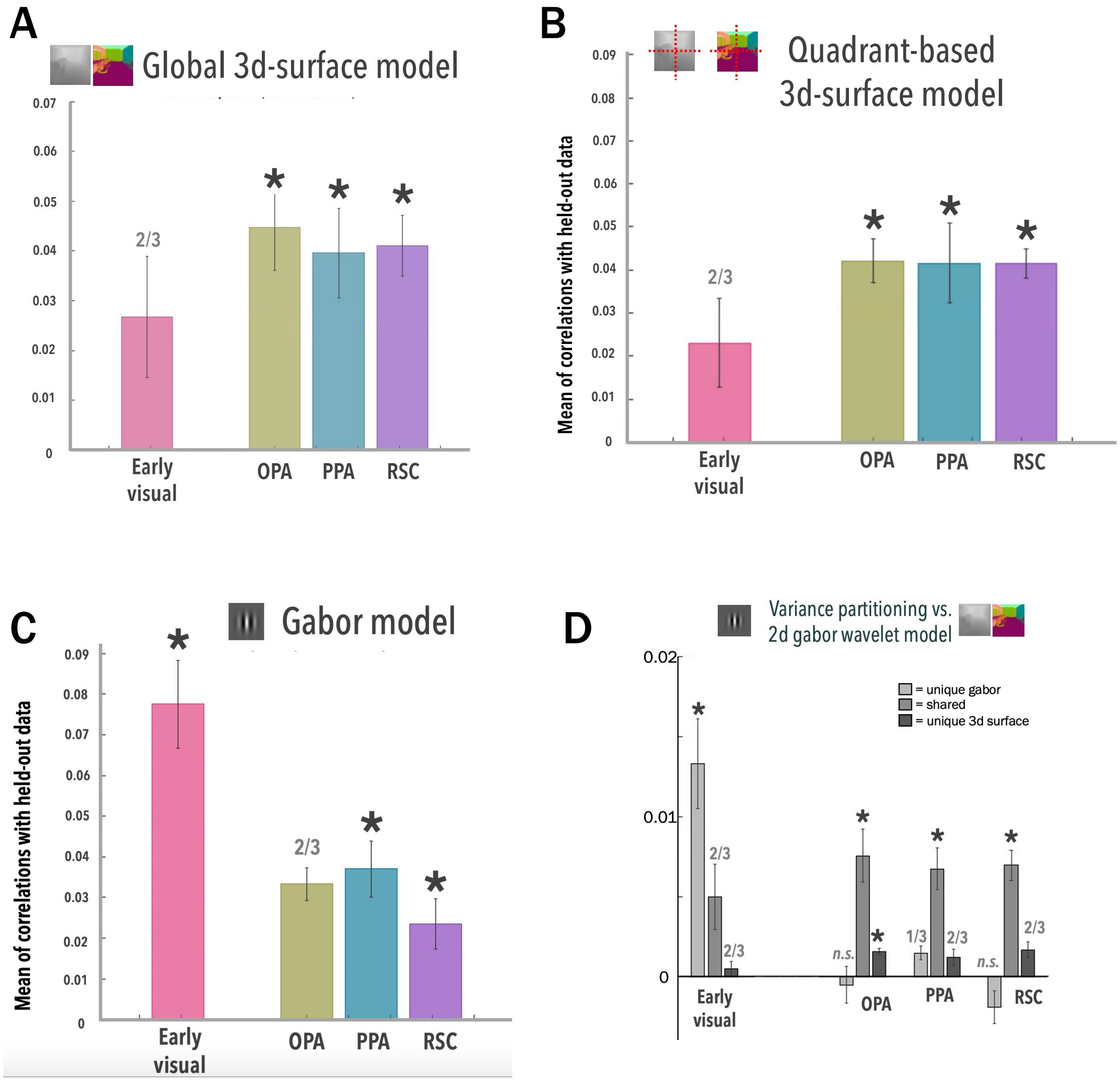


Figure 2.3. Results. Stars denote an effect that is significant in each of 3 participants; 2/3 denotes an effect that is significant for 2/3 participants, and n.s. denotes an effect not significant for any participants. (A) Prediction performance for the global 3d surface-based model. Performance is measured for each individual voxel for each cross-validation fold, and the measure for each ROI is the mean across folds, across voxels. (B) Prediction performance for the quadrant-based 3d surface model. (C) Prediction performance for the gabor model. (D) Variance partitioning for the global 3d surface-based model, vs. the gabor model. Y axis is proportion variance explained.

0.834, 0.0001) showing a very similar pattern of results to the global 3d model (Figure 2.3A). This may mean that both types of information are distributed similarly across all three scene-selective ROIs—however, future work is needed to test whether this pattern holds true even without sizeable correlations between feature sets.

Shared variance between 3d model and gabor wavelet model: As expected (Lescroart & Gallant, 2019), in early visual cortex, we found the largest amount of unique variance explained by the gabor model (all ps < 0.0001), while in scene-selective cortex there was little to no unique variance explained by the gabor model (OPA: ps < 0.72, 0.85, 0.70; PPA: (ps < 0.01, 0.06, 0.08); RSC: ps < 0.28, 0.23, 0.12). Compared to previous data with artificially generated scenes (Lescroart & Gallant, 2019), the pattern of results diverged slightly more for the global 3d model in scene-selective areas—in this data set, we found shared variance between the 2d and 3d models (OPA: all ps < 0.013; PPA: all ps < 0.0001; RSC: all ps < 0.0001) of a larger magnitude than unique 3d-model variance (Figure 2.3D; OPA: all ps < 0.0043; PPA: ps < 0.075, 0.0001, 0.0014; RSC: ps < 0.0006, 0.0001, 0.18). These results are unlikely to be explained by high correlations of 2d vs. 3d features across images in our stimulus set; there was a correlation of just 0.056 between these two feature sets' RDMs, corresponding to 0.31% variance. Instead, one explanation is the lack of real-world visual statistics in artificially generated images, compared to the real-world images used in this experiment—it may be the case that naturalistic 2d patterns of orientation and spatial frequency act as cues to 3d structure (Brady, Shafer-Skelton, & Alvarez, 2017) when present, and that breaking this relationship results in less unique 3d model variance. These results highlight the importance of using both highly controlled and naturalistic stimuli across experiments to gain a more complete picture of the visual system's representations.

Discussion

In this work we assessed the success of 3d surface information at predicting voxel responses to naturalistic scene images in scene-selective areas OPA, PPA, and RSC. We did this using Taskonomy (Zamir et al., 2019) deep neural networks to estimate distance and surface-direction maps used to compute the features for stimuli in the publicly available BOLD5000 data set (Chang et al., 2019). We found that 3d surface information significantly predicted responses in scene-selective areas, even in these naturalistic images. In most scene-selective regions and subjects, there was also significant 3D variance explained beyond the influence of 2D features. This information is detectable even using DNNs to estimate the distances and surface directions in photographs of real scenes, demonstrating that it is robust to visual properties of images and is not an artifact of artificially generated images used in previous work.

These results also show an interesting difference from previous findings. Surprisingly, shared variance between 2d and 3d features was notably larger in this experiment than in work using artificially generated movie stimuli (Lescroart & Gallant, 2019), highlighting the importance of studying visual representations by isolating different visual features of interest and by using stimulus sets that are ecologically valid in complementary ways. Because of the small number of repeated images in the BOLD5000 stimulus set, we were unable to calculate the noise-ceiling-normalized variance explained that would allow us to select voxels in the same way as Lescroart & Gallant (2019) or to compare absolute magnitudes of variance explained across these different experiments. Nevertheless, these results are apparent from the pattern of results across ROIs.

What types of behaviors might these 3d surface features support? We primarily focus on global 3d surface features here in order to relate our work to Lescroart & Gallant (2019)—these features are based on the proportions of stimulus pixels falling into each combination of distance and surface-direction bin across the entire image. This may capture information used to identify a new environment when entering or to recognize a new location across cuts when watching a movie—for example, openness is an important property that was captured by this model (Lescroart & Gallant, 2019) that can be used for categorization. However, other important global properties, like concavity (A. Cheng, Walther, Park, & Dilks, 2021), are not distinguishable without at least some coarse information about which pixels are where in the image. Thus we divided images into quadrants for the quadrant-based scene-surface model and computed the same features for each quadrant. This model should also pick up on information related to navigating a scene—for example, whether there are horizontal surfaces in the bottom half of the image. Given the differential roles of OPA vs. PPA in identification vs. navigation of scenes (Persichetti & Dilks, 2018), we thought we might find a differential pattern of these models across scene areas. Instead, we found a strikingly similar pattern across ROIs. An important question for future work is whether this was due to correlations between the two models' features in the stimulus set we used or due to similarly distributed information across areas.

What types of images may be suitable for this approach? It is notable that the images used to train the Taskonomy neural networks were almost exclusively indoor images (with some images including windows or extending to an outdoor patio, but ground-truth 3d annotations were either out-of-range or counted the pane of the window as the surface in the scene, even though they were visible in the image). Despite this, however, we still found mostly error-free distance and surface-normal maps, as well as significant prediction of scene-area responses using features based on these estimations. This may mean that the same principles the DNN learns to estimate 3d properties of indoor scenes translate to outdoor scenes as well. Regardless of the reason, this result is a demonstration of the flexibility of this approach. Even if there may be images with less-good feature estimations, euclidean-distance and surface-normal estimation are relatively common computer vision tasks, and the basic approach of estimating neural activity using these features could be extended to future DNNs trained on different types of images. We believe the approach we used here is a powerful tool to make use of existing data sets to help close the gap between our knowledge about representations of discrete visual objects and of the 3-dimensional spaces we interact with every day.

Work with other quantifications of 3d scene information: One similar approach to the current work uses DNN features themselves as predictors for an encoding model (Wang, Wehbe, & Tarr, n.d.). The authors categorize different types of neural networks into those trained to do 2d or 3d tasks, finding that latent-space activations of 3d DNNs predict human voxel responses in scene-selective areas better than those of 2d DNNs. While we don't have a precise understanding of the type of representation contained in the DNN activations above, the present work uses human-interpretable spatial features that give a more specific picture of the type of 3d representations present.

Other work (Bonner & Epstein, 2017; Greene & Oliva, 2009) has estimated spatial features via individually drawn annotations. This process takes time and potentially money, and researchers are limited to quantifying properties that are easy for participants to report on a computer. We hope the present work is a convincing demonstration that DNNs can be used to estimate more detailed properties that may otherwise be impossible or very difficult to collect from human data.

Conclusion

This work was motivated by the relative dearth of knowledge the field has about visual processing of complex features like the 3d spatial structure of scenes, compared to representations of lower-level features or discrete objects (Brady et al., 2019). We found that even in naturalistic images, both global and quadrant-based 3d scene-surface representations can predict voxel responses in scene-selective cortex. This work highlights an exciting new possibility for investigating the format of visual representations of 3d scene surfaces, a topic that has had a high barrier to entry in the past. We also provide a proof of concept for using DNNs to estimate ground-truth 3d information from publicly available fMRI data sets that would otherwise be unusable for this purpose—this is an exciting new technique that lowers the barrier to entry for studying 3d configurations of surfaces in scenes, a critical piece of information to our visual systems as we move through the world every day.

Acknowledgements

Chapter 2 will be prepared to submit for publication. Shafer-Skelton, A.; Brady, Timothy F.; Serences, John T. The dissertation was the primary investigator and author of this material.

References

Alvarez, G. a, & Oliva, A. (2009). 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, 106(18), 7345–7350. https://doi.org/10.1073/pnas.0808981106

Bonner, M. F., & Epstein, R. A. (2017). Coding of navigational affordances in the human visual system. Proceedings of the National Academy of Sciences, 114(18), 4793–4798. https://doi.org/10.1073/pnas.1618228114

Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. Journal of Experimental Psychology: Human Perception and Performance, 5, 0–17. https://doi.org/10.1037/xhp0000399

Brady, T. F., Störmer, V. S., Shafer-Skelton, A., Williams, J. R., Chapman, A. F., & Schill, H. M. (2019). Scaling up visual attention and visual working memory to the real world. Psychology of Learning and Motivation - Advances in Research and Theory, 70, 29–69. https://doi.org/10.1016/bs.plm.2019.03.001

Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., & Aminoff, E. M. (2019). BOLD5000, a public fMRI dataset while viewing 5000 visual images. Scientific Data, 6(1), 49. https://doi.org/10.1038/s41597-019-0052-3

Cheng, A., Walther, D. B., Park, S., & Dilks, D. D. (2021). Concavity as a diagnostic feature of visual scenes. NeuroImage, 232, 117920. https://doi.org/10.1016/j.neuroimage.2021.117920

Cheng, K. (1986). A purely geometric module in the rat's spatial representation*. Cognition, 23, 149–178.

Dilks, D. D., Julian, J. B., Paunov, A. M., & Kanwisher, N. (2013). The occipital place area is causally and selectively involved in scene perception. Journal of Neuroscience, 33(4), 1331–1336. https://doi.org/10.1523/JNEUROSCI.4081-12.2013

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. Nature, 392(6676), 598–601. https://doi.org/10.1038/33402

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual Scenes Are Categorized by Function. Journal of Experimental Psychology: General. https://doi.org/10.1037/xge0000129.supp

Greene, M. R., & Hansen, B. C. (2020). Disentangling the Independent Contributions of Visual and Conceptual Features to the Spatiotemporal Dynamics of Scene Categorization. The Journal of Neuroscience, (April), JN-RM-2088-19. https://doi.org/10.1523/jneurosci.2088-19.2020

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: seeing the forest without representing the trees. Cognitive Psychology, 58(2), 137–176. https://doi.org/10.1016/j.cogpsych.2008.06.001

Groen, I. I. A., Ghebreab, S., Lamme, V. A. F., & Scholte, H. S. (2016). The time course of natural scene perception with reduced attention. Journal of Neurophysiology, 115(2), 931–946. https://doi.org/10.1152/jn.00896.2015

Groen, I. I. A., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. ELife, 7. https://doi.org/10.7554/eLife.32962

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

Hermer, L., & Spelke, E. S. (1994). A geometric process for spatial reorientation in young children. Nature, 370(6484), 57–59. https://doi.org/10.1038/370057a0

Julian, J. B., Ryan, J., Hamilton, R. H., & Epstein, R. A. (2016). The Occipital Place Area Is Causally Involved in Representing Environmental Boundaries during Navigation. Current Biology, 26(8), 1104–1109. https://doi.org/10.1016/j.cub.2016.02.066

Landau, B., & Lakusta, L. (2009). Spatial representation across species: geometry, language, and maps. Current Opinion in Neurobiology, 19(1), 12–19. https://doi.org/10.1016/j.conb.2009.02.001

Lescroart, M. D., & Gallant, J. L. (2019). Human Scene-Selective Areas Represent 3D Configurations of Surfaces. Neuron, 101(1), 178–192. https://doi.org/10.1016/j.neuron.2018.11.004

Martin Cichy, R., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. NeuroImage, 153, 346–358. https://doi.org/10.1016/j.neuroimage.2016.03.063

Mullin, C. R., & Steeves, J. K. E. (2011). TMS to the Lateral Occipital Cortex Disrupts Object Processing but Facilitates Scene Processing.

Mullin, C. R., & Steeves, J. K. E. (2013). Consecutive TMS-fMRI Reveals an Inverse Relationship in BOLD Signal between Object and Scene Processing. https://doi.org/10.1523/JNEUROSCI.2537-13.2013

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. Progress in Brain Research, 155.

Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. The Journal of Neuroscience, 31(4), 1333–1340. https://doi.org/10.1523/JNEUROSCI.3885-10.2011

Persichetti, A. S., & Dilks, D. D. (2018). Dissociable Neural Systems for Recognizing Places and Navigating through Them. The Journal of Neuroscience, 38(48), 10295–10304. https://doi.org/10.1523/jneurosci.1200-18.2018

Sax, A., Emi, B., Zamir, A. R., Guibas, L., Savarese, S., & Malik, J. (2018). Mid-Level Visual Representations Improve Generalization and Sample Efficiency for Learning Visuomotor Policies. Retrieved from http://arxiv.org/abs/1812.11971

Silson, E. H., Steel, A., Kidder, A., Gilmore, A. W., & Baker, C. I. (2019). Distinct subdivisions of human medial parietal cortex support recollection of people and places. ELife, 8. https://doi.org/10.7554/eLife.47391

Wang, A. Y., Wehbe, L., & Tarr, M. J. (n.d.). Neural Taskonomy: Inferring the Similarity of Task-Derived Representations from Brain Activity, (NeurIPS), 708016. https://doi.org/10.1101/708016

Zamir, A., Sax, A., Shen, W., Guibas, L., Malik, J., & Savarese, S. (2019). Taskonomy: Disentangling task transfer learning. In IJCAI International Joint Conference on Artificial Intelligence (Vol. 2019-Augus, pp. 6241–6245). https://doi.org/10.24963/ijcai.2019/871

CHAPTER 3: Stimulus dependence of 3D-scene-surface representations in scene-selective cortex

Abstract

Recent analysis techniques allow us to better quantify and compare different types of complex visual information in natural scene stimuli. Using two different tasks—one more dependent on scene structure and one less—we investigated representations of the 3D geometry of natural scenes. Specifically, are scene-selective cortical areas sensitive to *global* 3D properties that may aid in scene identification and re-orientation in everyday life, and/or *spatially specific* configurations of 3D-oriented surfaces that may be more useful for acting on and moving through our environment? By iteratively sampling potential stimuli from the Taskonomy training set, we created a stimulus set in which global and spatially specific features summarizing the 3D configurations of surfaces were more separable than in previous work. This process also resulted in real-world photographs that, as a set, separate 2D gabor-wavelet features from 3D scene-surface features better than in previous studies. This allowed us to follow up on findings in Chapter 2 that real-world scene photographs elicited more *shared* 2D/3D information (not uniquely attributable to either) compared to unique-3D information than in Lescroart & Gallant (2019). Surprisingly, with this information more separable across stimuli, we find evidence for more 2D than 3D information in scene-selective cortex, suggesting previous findings of greater 3D information could have been due to computer-generated stimuli that render naturalistic 2D depth cues less useful (Lescroart & Gallant, 2019), and/or to our use of static photographs that cannot provide motion-parallax as a possible depth cue. These findings highlight the importance of investigating complex visual information in natural scenes using a variety of stimuli that isolate different aspects of the natural world.

Introduction

A large body of evidence in vision science argues for a fascinating division: that representations of extended visual scenes are distinct from those of discrete visual objects (e.g., Dilks, Julian, Paunov, & Kanwisher, 2013; Epstein & Kanwisher, 1998). While early work suggests that part of what makes scene representations special is information about configurations of major surfaces like walls and floors (e.g., Epstein & Kanwisher, 1998; Hermer & Spelke, 1994; Landau & Lakusta, 2009), less is known about the precise format of this information or how such a distinction may impact cognitive processes such as visual attention and memory. This is challenging to study with behavioral paradigms, in part because of the large amount of information needed to characterize these representations well enough to disambiguate them from those of global properties (like "openness" or "navigability") and low-level features (like orientation or spatial-frequency information). fMRI work, however, is well positioned to study these types of representations, since their presence can be assessed without the difficulty of finding a reporting method that captures enough information. Three well-studied regions of cortex offer fruitful areas to target: 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 there is already some evidence that they contain information relevant to us as we navigate through the 3D world (Bonner & Epstein, 2017; Henriksson, Mur, & Kriegeskorte, 2019; Park & Park, 2020), including causal evidence in OPA (Dilks et al., 2013).

One study in particular, Lescroart & Gallant (2019), more precisely examines the format of this information, finding evidence for the presence of global summaries of information about the 3D configurations of major surfaces in scenes, summarized across entire images. These 3D features predict voxel responses beyond 2D patterns of orientation and spatial frequency, offering evidence of a scene-specific representation. This global, whole-image summary information may be useful for re-orienting when entering a new space—for example, going through a doorway and detecting a left-facing and right-facing wall may signal to us that we're in a hallway. However, without information about where in our visual field those surfaces are, we would fail to effectively navigate to where we need to go. Thus, the present work aimed to test the presence of *spatially specific* 3D scene surface representations, which may support spatial navigation and visual exploration beyond scene identification. In particular, using fMRI encoding models, we aimed to look at the extent to which voxels in scene-responsive regions of visual cortex contain information about the global spatial structure of the surfaces in an image, as in Lescroart and Gallant (2019), as well as the (local) spatial structure of surfaces in each quadrant of the image. To do so, we designed a stimulus set that allowed these two to be distinguished, by making them relatively un-correlated.

We also aimed to further investigate the relationship between 3D scene-surface information and 2D orientation and spatial-frequency information, since the stimuli we chose to differentiate between global vs. spatially specific 3D visual features also turned out to have the potential to differentiate between 3D and 2D models. In addition, since Chapter 2 used *estimated* 3D features to probe 3D representations, the present experiment offers an important opportunity to verify this finding in a different stimulus set, in this case, with images used to train the Taskonomy DNN model. This helps ensure the generalizability of the results from Chapter 2 and reduces the dependence on the Taskonomy DNNs' (Zamir et al., 2019) depth and surface-direction estimates in novel images. Further, while the BOLD5000 images (Chang et al., 2019) in Chapter 2 were taken from the internet, and many appear to be carefully composed images taken to be visually appealing, the images in the Taskonomy training set used here were taken to teach DNNs about visual scenes we encounter on a day-to-day basis and appear much more mundane and potentially more representative of our day-to-day experience. Our use of both types of naturalistic scene photographs offers an important complement to the 3D-rendered stimuli used in Lescroart & Gallant (2019) and Henriksson, Mur, & Kriegeskorte (2019), preserving more naturalistic texture patterns that might influence our processing of 3D surfaces in our everyday lives.

In the present work, we use voxelwise encoding models to quantify the strength of "information" about different visual features. While we manipulated the relevance of the 3D spatial structure of the scene, as well as the spatial extent of the likely locations of probe shapes (see *Methods* for details), we don't focus on those manipulations in this paper. First, we compare two types of 3D information: (1) global (whole-image) summaries of 3D configurations of the major surfaces in scenes (as in Lescroart & Gallant, 2019), which may capture global properties such as openness, and (2) spatially specific summaries (based on image quadrant) that also capture where these surfaces are *within* the image, information critical to guiding our actions within the world. In addition, our real-world photograph stimuli (as contrasted with the rendered scenes used in previous work) preserve texture information *within* images that could influence our visual processing in real life—meanwhile, by selecting our stimuli purposefully, we have ensured that *across* images, feature sets of interest are more separable than in previous work, affording a prime opportunity to re-examine the relationships of each of the above types of 3D scene information to (3) gabor wavelet features that capture 2D patterns of orientation and spatial frequency in each image. By ensuring that the images with the most similar 2D features are not the images with the most similar 3D features, we allow our model fits to more uniquely pick up on the features that voxels are most responsive to.

Methods

Participants: 8 participants (6 female) 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 three fMRI sessions lasting approximately 2 hours each. 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).

Participants completed one retinotopic mapping session with the same scan parameters as functional data in the above experimental sessions. 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 (Dale, Fischl, & Sereno, 1999) 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 (see *Regions of Interest*). We used Freesurfer's manual and automatic boundary-based registration tools (Greve & Fischl, 2009) to generate transformation matrices that were then used by FSL FLIRT (Jenkinson, Bannister, Brady, & Smith, 2002; Jenkinson & Smith, 2001) 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 (Jenkinson et al., 2002), 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 signal 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.

Regions of interest: For regions V1, V2, and V3, we followed previously established retinotopic mapping protocols (Engel, Glover, & Wandell, 1997; Sereno et al., 1995; Wandell, Dumoulin, & Brewer, 2007). 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, using a false discovery cutoff of $q < 0.05$. For regions OPA, PPA, and MPA/RSC, we used a combination of localizer runs from past experiments and the present experiment (stimuli from Grill-Spector lab), for a total of 4-10 functional localizer runs per subject. 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 S3.4 for main analysis results as a function of number of voxels included). This participant is omitted in MPA/RSC analyses.

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. 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). Order of mini-blocks within runs 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. 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.

Stimuli and Tasks: 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 (Sanocki, 2003; Shafer-Skelton & Brady, 2019). In *shape judgment* runs, the participant was instructed to respond whether the left or right shape was a square (/diamond). 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 S3.1 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.

Stimulus selection: We include 300 stimulus images from the Taskonomy training set selected to maximally differentiate between spatial and non-spatial 3d 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 (2017), 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 global and quadrant-based feature RDMs, similar to the correlation between these two RDMs in the BOLD5000 stimuli used in Chapter 2. 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 global and 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 (global 3D vs. gabor: -0.002) than both Chapter 2's stimuli (0.056) and Lescroart et al.'s (2019) validation data set (0.274). This pattern was robust to choice of distance metric (see Table S3.1).

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 the narrow, medium, and broad sets of dot pairs didn’t differ substantially in depth difference between dots.

Feature sets: 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 similar to Lescroart & Gallant (2019), achieving similar results as with the 300-feature model. These features were generated using code from https://github.com/ gallantlab/motion_energy_matlab.

G*lobal* and *quadrant-based 3d scene surface features* were computed in much the same way as in Chapter 2. This time, we used ground-truth distance and surface-normal (surface direction) image maps that were used as the training set for the Taskonomy (Zamir et al., 2019) neural networks. To generate features from these maps, we adapted code from Lescroart & Gallant (2019), first grouping distances into 3 bins (vs. 10 in Lescroart & Gallant, 2019). We included the same 9 surface-direction bins contained in Lescroart & Gallant (2019). In the *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 *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 fitting 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 the data in the fitting portion of that main fold. There were 13 possible ridge parameters: 0, as well as 12 values logarithmically spaced between 10^{-2} and 10^{5}. Early analyses from Chapter 2 showed that this ridge parameter selection matched performance using a more extensive set of ridge parameter choices.

Similarly to Chapter 2, 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 a given 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.

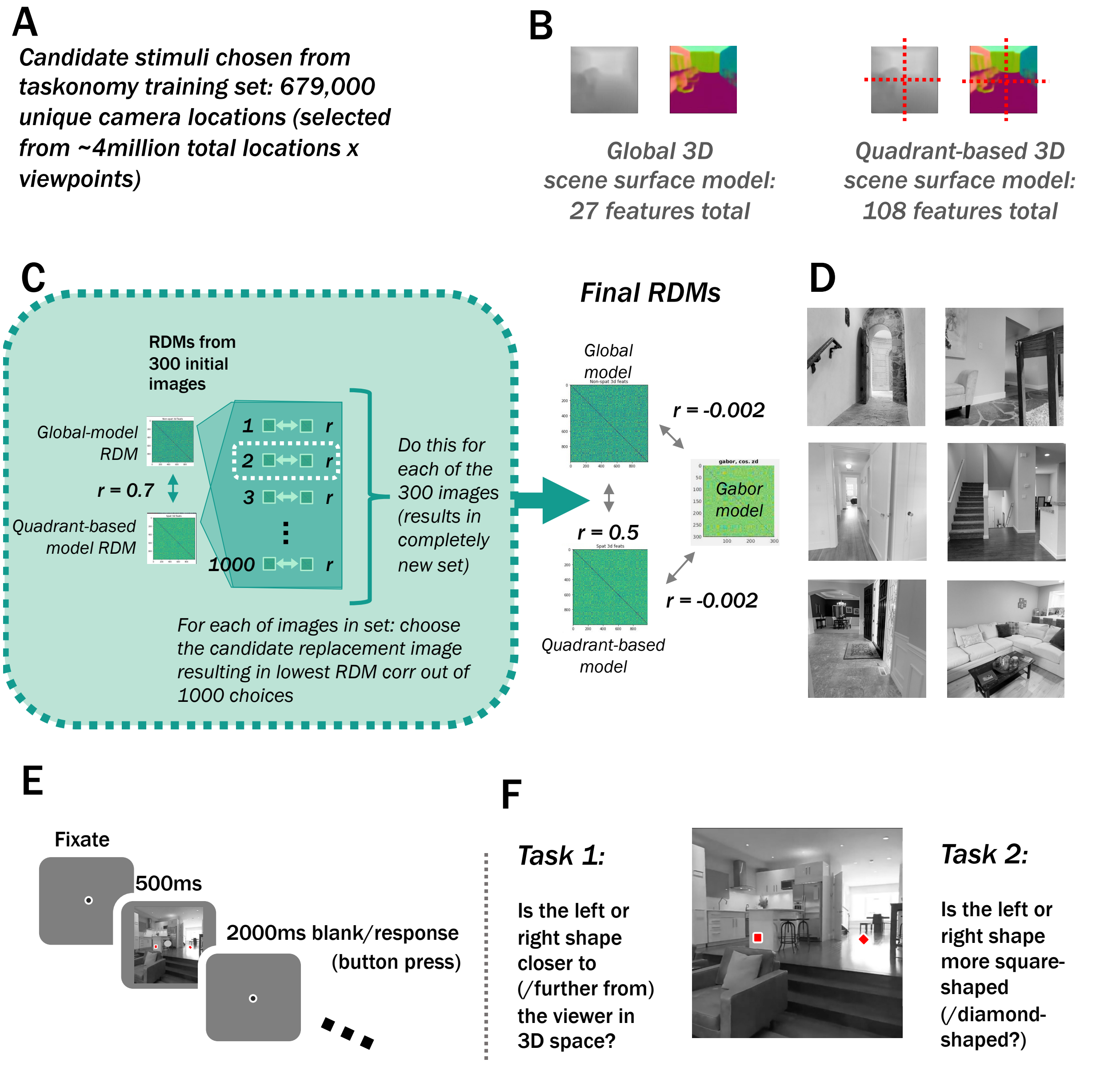

A
Candidate stimuli chosen from taskonomy training set: 679,000 unique camera locations (selected from ~4million total locations x viewpoints)
B
Global 3D scene surface model: 27 features total
Quadrant-based 3D scene surface model: 108 features total
C
RDMs from 300 initial images
Global-model RDM
r = 0.7
Quadrant-based model RDM
1 r
2 r
3 r
1000 r
Do this for each of the 300 images (results in completely new set)
For each of images in set: choose the candidate replacement image resulting in lowest RDM corr out of 1000 choices
Final RDMs
Global model
r = -0.002
Gabor model
r = 0.5
r = -0.002
Quadrant-based model
D
E
Fixate
500ms
2000ms blank/response (button press)
F
Task 1:
Is the left or right shape closer to (/further from) the viewer in 3D space?
Task 2:
Is the left or right shape more square-shaped (/diamond-shaped?)

Figure 3.1. 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 (Zamir et al., 2019) training set. (C) The 300 stimulus images were chosen to minimize spatial- vs. quadrant-based-model RDMs. An initial set of stimulus images (chosen similarly to Groen et al., (2007)) 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 3.2 shows the fit of each model in each visual region.

EVC performance serves as a baseline, replicating previous results: As expected, we find significant gabor- and 3D-surface-model performance in early visual areas (all *p*s < 0.0002, uncorrected; Table S3.2A), showing a similar pattern as in previous work, as well as similar magnitudes of cross-validated prediction performance (Figures 3.B in both the present chapter and in Lescroart & Gallant, 2019).

Stronger 2D vs. 3D representations in scene areas; no evidence for uniquely 3D representations: In scene areas, we had intended to use gabor performance only as a baseline as well, since previous work (Chapter 2; Henriksson et al., 2019; Lescroart & Gallant, 2019) had converged on the presence of unique 3D scene-surface representations. While we found significant prediction performance for all three models in scene-selective cortex (all *p*s < 0.014; see Figure 3.2A and Table S3.2A), the lower 3D-scene-surface performance compared to 2D gabor-wavelet performance (Figure 3.2A; all ps < 0.0025) prompted us to attempt to replicate the previous 2D-vs-3D variance-partitioning results (Lescroart & Gallant, 2019) in these areas—surprisingly, we found *only* consistent evidence of unique *2D* representations (vs. quadrant-based: OPA: p = 0.0006; PPA: p = 0.0002; RSC: p = 0.026; vs. global: OPA: 0.0018; PPA: 0.005; RSC: 0.053), as well as *shared* 2D/3D-quadrant-based (OPA: p = 0.0019; PPA: p = 0.0004; RSC: p = 0.026) and shared 2D/3D-global (OPA: 0.0069; PPA: 0.0018; RSC: 0.0048) representations, with no evidence for uniquely 3D representations (global 3D model: all *p*s > 0.13; quadrant-based 3D model: all *p*s > 0.41; Figure 3.2B and C and Table S3.2 B and C). Note that RSC results follow the same pattern but are not as robust as OPA/PPA results, consistent with previous work. In OPA and PPA, the unique variance explained by the gabor model was significantly greater than either the quadrant-based-3D-model (ps < 0.0008) or the global-3D-model (ps < 0.004), as well as than variance shared with the quadrant-based model (ps < 0.0027) or global model (ps < 0.0061). Again, RSC showed a similar but less reliable pattern (Figure 3.2 B and C). This was the opposite pattern of results as found in Lescroart & Gallant (2019).

Next, we investigated whether this difference could be explained by methodological differences. Plotting each model's performance as a function of the number of voxels included in each ROI, we find a strikingly consistent pattern of relative model performance, arguing against this pattern being an artifact of different voxel selection procedures across studies (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, our main conclusions hinge on the *pattern* irrespective of magnitude. Next, a sliding window analysis (Figure S3.3) shows that these results are also consistent across timepoints. We also show that this pattern of results was not due to collapsing across scene-task and orthogonal-task data—the scene-task data alone show an almost identical pattern of model performances in all ROIs (Figure S3.2).

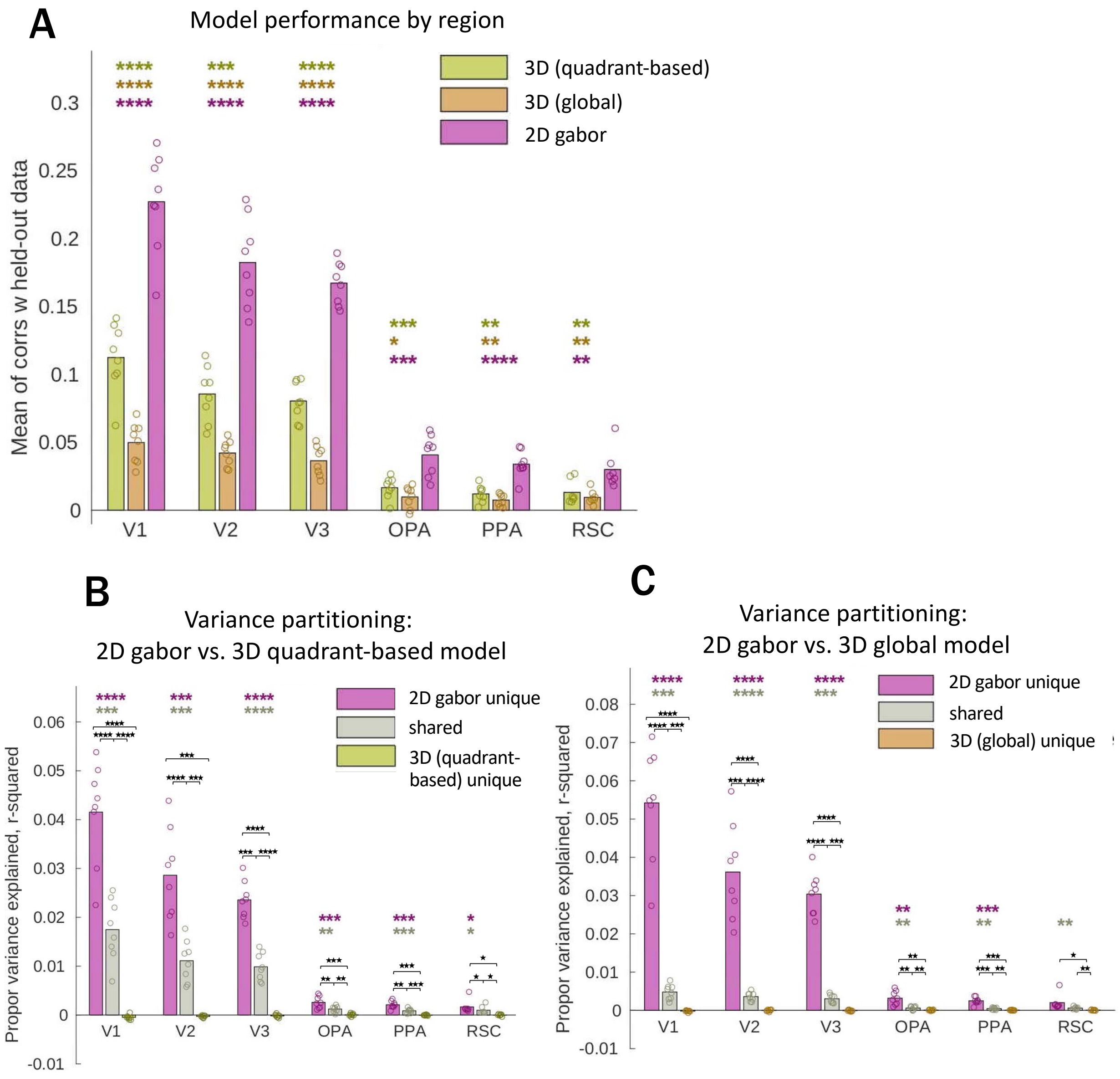


Figure 3.2. Dots represent individual participant data. Stars denote significant two-tailed effects in the positive direction (**** = p<0.0001; *** = p<0.001; ** = p<0.01; * = p<0.05, uncorrected). Colored stars correspond to individual tests, and brackets indicate paired tests. Note that magnitudes across ROIs are not interpretable, but patterns within ROIs are. (A) Results collapsed across all task data. (B) Variance partitioning results for 2D gabor model vs. 3D quadrant-based 3D model. (C) Variance partitioning results for 2D gabor model vs. 3D global model.

Discussion

This project investigated the neural representation of visual scenes in scene-responsive visual cortex using natural scene photograph stimuli selected to best separate *global* and *spatially specific* (quadrant-based) representations of 3D scene surfaces. After the stimulus selection procedure, each of the 3D models was also more separable from a 2D gabor wavelet model than in previous work, which gave us an opportunity to also distinguish 2D vs 3D representations in general. Contrary to previous work, we find higher prediction performance for a 2D-gabor-wavelet model compared to 3D-scene-surface models, even in scene-selective cortex. We also find evidence for uniquely 2D representations but not uniquely 3D representations in scene-selective areas OPA, PPA, and MPA/RSC, contrary to the dominant claim about OPA in particular (Bonner & Epstein, 2017; Henriksson et al., 2019; Lescroart & Gallant, 2019; Park & Park, 2020).

No evidence for representations uniquely attributable to 3D scene surface information: Our variance partitioning analysis finds no evidence that 3D scene surface features can uniquely explain activity in scene-selective cortex—these results conflict most directly with Lescroart & Gallant (2019), who used similar analysis techniques but different stimuli. Two differences in our stimuli could explain this: (1) in contrast to work with computer-generated stimuli (Henriksson et al., 2019; Lescroart & Gallant, 2019), which showed 3D scene surface features providing an added benefit beyond 2D models, our stimuli consisted of naturalistic photographs sampled from the real world. That means that *within* each photograph, naturalistic textures and relationships among different types of information were closer to in real life, which in our study may have encouraged participants' visual systems to rely more on 2D cues to infer

3D information. (2) *Across* images in our stimulus set, 2D and 3D features are more separable than in both Ch. 2 and in Lescroart & Gallant (2019; see Table S3.1 for comparison), positioning us to be better able to differentiate between shared 2D/3D variance and variance uniquely explained by 2D or 3D features.

Note that our results motivate further investigation of one other stimulus difference: Lescroart & Gallant's stimuli were 3D-rendered videos and thus may have provided participants with one other monocular depth cue to rely on: motion parallax. Future work could model this information to determine its relative contributions to 3D stimulus representations. If it turned out to be a major contributor to these differing results, it could further incentivize a shift away from static visual stimuli and towards using naturalistic movies.

However, while stimulus differences are a possible cause, differences between the present work and Lescroart & Gallant (2019) can't be straightforwardly explained by analysis differences. First, when varying ROI size (even with different numbers across different regions), the pattern of results is very robust, even within individual participants (Figure S3.4). While we use a slightly shorter time window than Lescroart & Gallant (2019) and the BOLD5000 data in Chapter 2, our pattern of results is consistent across a wide range of timepoints (Figure S3.3). Finally, we see the same pattern for the scene-relevant task (Figure S3.2) as we do collapsed across tasks (main results). It is also worth noting that both the present study and Lescroart & Gallant (2019) fit models using cross-validation, making the results robust to overfitting.

There are also differences between the present results and *Chapter 2*'s results: first, the large amount of shared variance in Chapter 2, compared to the large amount of unique-2D variance in the current data. This difference may be due to 2D vs. 3D features being more separable across images in the present feature set than in Chapter 2. For example, with these features being more separable across images, variance that had previously not been uniquely attributable to either model may, in Chapter 3, have become uniquely attributable to the gabor wavelet model. A second difference is that there appeared to be a small amount of unique-3D information in Chapter 2. While it was not significant in every participant/ROI, it appeared to at least show a relatively consistent pattern—meanwhile, we found no detectable 3D information in the present data. Given this difference, other publicly available fMRI data sets, such as the Natural Scenes Dataset (Allen et al., 2021) may serve to add more data points, helping to more precisely estimate the strength of any unique-3D information using the methods in Chapter 2.

Other studies that explicitly compared 3D to 2D information have found compatible results with ours, even when interpretations are slightly different. First, while investigating local-scene-affordance information, a related type of 3D visual information, Bonner & Epstein (2017) find a similar pattern of higher performance for 2D features (a gist model) than their 3D information. The question of whether there is unique 3D information in their data hinges on whether it's reasonable to discount 2D variance that's redundant with information in EVC. It's also worth noting that they find more 2D gist than 3D information even though Lescroart & Gallant (2019) find that a 2D gist model performs less than half as well as a 2D gabor wavelet model. This model difference also potentially affects Henriksson et al. (2019), which used a gist model as their measure of 2D information as well.

In summary, differences from previous results/conclusions may arise from: artificially generated scenes breaking the correspondence of 2D cues to the 3D structure of the world (Lescroart & Gallant, 2019; Henriksson et al., 2019), the presence of one other monocular depth cue (Lescroart & Gallant, 2019), 2D vs. 3D features sets that were not entirely separable (Ch. 2;

Lescroart & Gallant, 2019), and/or choice of a lower-performing 2D feature set in previous work (Henriksson et al., 2019, Figure 7A; c.f. Figure 3B in Lescroart & Gallant, 2019).

Implications for the visual system: The present pattern of results, considered in conjunction with other data, suggests that, when 2D orientation info that can be used as a cue to the 3D geometry of surfaces, there is less (or no) detectable unique 3D surface information in scene-selective cortex. This converges with evidence that 2D orientations/spatial frequencies are computationally sufficient to infer global scene properties related to 3D layout (Ross & Oliva, 2010) and that human behavior does indeed seem to rely on these features in scene processing tasks (Brady, Shafer-Skelton, & Alvarez, 2017). In summary, the present work emphasizes that, for naturalistic photographs, a possible underpinning of cortical scene-selective areas' 3D representations is via 2D patterns of orientation and spatial frequency, and that separating abstract representations of 3D surfaces from such 2D patterns may be difficult in natural scenes, where they are systematically related.

In this work, we also manipulated task, as well as the likely spatial regions of task-relevant shapes. In particular, we originally sought to gain insight into a special characteristic of scene representations: the reduced attention with which some types of scene information can apparently be processed (Alvarez & Oliva, 2009; Groen, Ghebreab, Lamme, & Scholte, 2016). If we had found unique-3D information in scene-selective cortex, that information could have served as a marker of scene-specificity, and we could have investigated its robustness to our task manipulation. While the patterns of orientation and spatial frequency captured by the gabor model may indeed be the basis of scene-specific representations, they are not themselves diagnostic and so do not lend themselves to understanding this task effect. Future work could address this by more directly quantifying relationships between this information and scene-specific behaviors it may support.

Conclusion: The present work further examines the presence of 2D and 3D representations in scene-selective cortex while participants viewed naturalistic scene-photograph stimuli. In contrast to Lescroart & Gallant (2019), we find that, using a stimulus set in which 3D-scene-surface and 2D-gabor information are also more separable across images than in previous work, we find a marked reverse pattern of greater 2D vs. 3D information in scene-selective cortex, as well as no detectable evidence of 3D information. This emphasizes the importance of investigating complex types of visual information using complementary stimulus sets that preserve different aspects of the natural world.

Acknowledgements:

Chapter 3 will be prepared to submit for publication. Shafer-Skelton, A.; Brady, Timothy F.; Serences, John T. The dissertation was the primary investigator and author of this material.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Dowdle, L. T., Caron, B., … Kay, K. (2021). A massive 7T fMRI dataset to bridge cognitive and computational neuroscience. *BioRxiv*, 2021.02.22.432340. https://doi.org/10.1101/2021.02.22.432340

Alvarez, G. a, & Oliva, A. (2009). 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*, *106*(18), 7345–7350. https://doi.org/10.1073/pnas.0808981106

Bonner, M. F., & Epstein, R. A. (2017). Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences*, *114*(18), 4793–4798. https://doi.org/10.1073/pnas.1618228114

Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 0–17. https://doi.org/10.1037/xhp0000399

Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., & Aminoff, E. M. (2019). BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, *6*(1), 49. https://doi.org/10.1038/s41597-019-0052-3

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis. *NeuroImage*, *194*, 179–194.

Dilks, D. D., Julian, J. B., Paunov, A. M., & Kanwisher, N. (2013). The occipital place area is causally and selectively involved in scene perception. *Journal of Neuroscience*, *33*(4), 1331–1336. https://doi.org/10.1523/JNEUROSCI.4081-12.2013

Engel, S. A., Glover, G. H., & Wandell, B. A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cerebral Cortex*, (7), 181–192.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601. https://doi.org/10.1038/33402

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2017). 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*, 207530. https://doi.org/10.1101/207530

Henriksson, L., Mur, M., & Kriegeskorte, N. (2019). Rapid invariant encoding of scene layout in human OPA. *BioRxiv*, 577064. https://doi.org/10.1101/577064

Hermer, L., & Spelke, E. S. (1994). A geometric process for spatial reorientation in young children. *Nature*, *370*(6484), 57–59. https://doi.org/10.1038/370057a0

Iris, X., Groen, I. A., Ghebreab, S., Lamme, V. A. F., & Scholte, H. S. (2016). The time course of natural scene perception with reduced attention. *Journal of Neurophysiology*, *115*, 931–946. https://doi.org/10.1152/jn.00896.2015.-Attention

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1006/nimg.2002.1132

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*(2), 143–156. https://doi.org/10.1016/S1361-8415(01)00036-6

Landau, B., & Lakusta, L. (2009). Spatial representation across species: geometry, language, and maps. *Current Opinion in Neurobiology*, *19*(1), 12–19. https://doi.org/10.1016/j.conb.2009.02.001

Lescroart, M. D., & Gallant, J. L. (2019). Human Scene-Selective Areas Represent 3D Configurations of Surfaces. *Neuron*, *101*(1), 178–192. https://doi.org/10.1016/j.neuron.2018.11.004

Park, J., & Park, S. (2020). Coding of navigational distance and functional constraint of boundaries in the human scene-selective cortex. *Journal of Neuroscience*, *40*(18), 3621–3630. https://doi.org/10.1523/JNEUROSCI.1991-19.2020

Ross, M. G., & Oliva, A. (2010). Estimating perception of scene layout properties from global image features. *Journal of Vision*, *10*(1), 2.1-25. https://doi.org/10.1167/10.1.2

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

Sereno, A. M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, T. J., … Tootell, R. B. H. (1995). Borders of Multiple Visual Areas in Humans Revealed by Functional Magnetic Resonance Imaging. *Science*, *268*(5212), 889–893.

Shafer-Skelton, A., & Brady, T. F. (2019). Scene layout priming relies primarily on low-level features rather than scene layout. *Journal of Vision*, *19*(1), 1–33. https://doi.org/10.1167/19.1.14

Wandell, B. A., Dumoulin, S. O., & Brewer, A. A. (2007). Visual field maps in human cortex. *Neuron*, *56*(2), 366–383. https://doi.org/10.1016/j.neuron.2007.10.012

Zamir, A., Sax, A., Shen, W., Guibas, L., Malik, J., & Savarese, S. (2019). Taskonomy: Disentangling task transfer learning. In *IJCAI International Joint Conference on Artificial Intelligence* (Vol. 2019-Augus, pp. 6241–6245). https://doi.org/10.24963/ijcai.2019/871

Supplement:

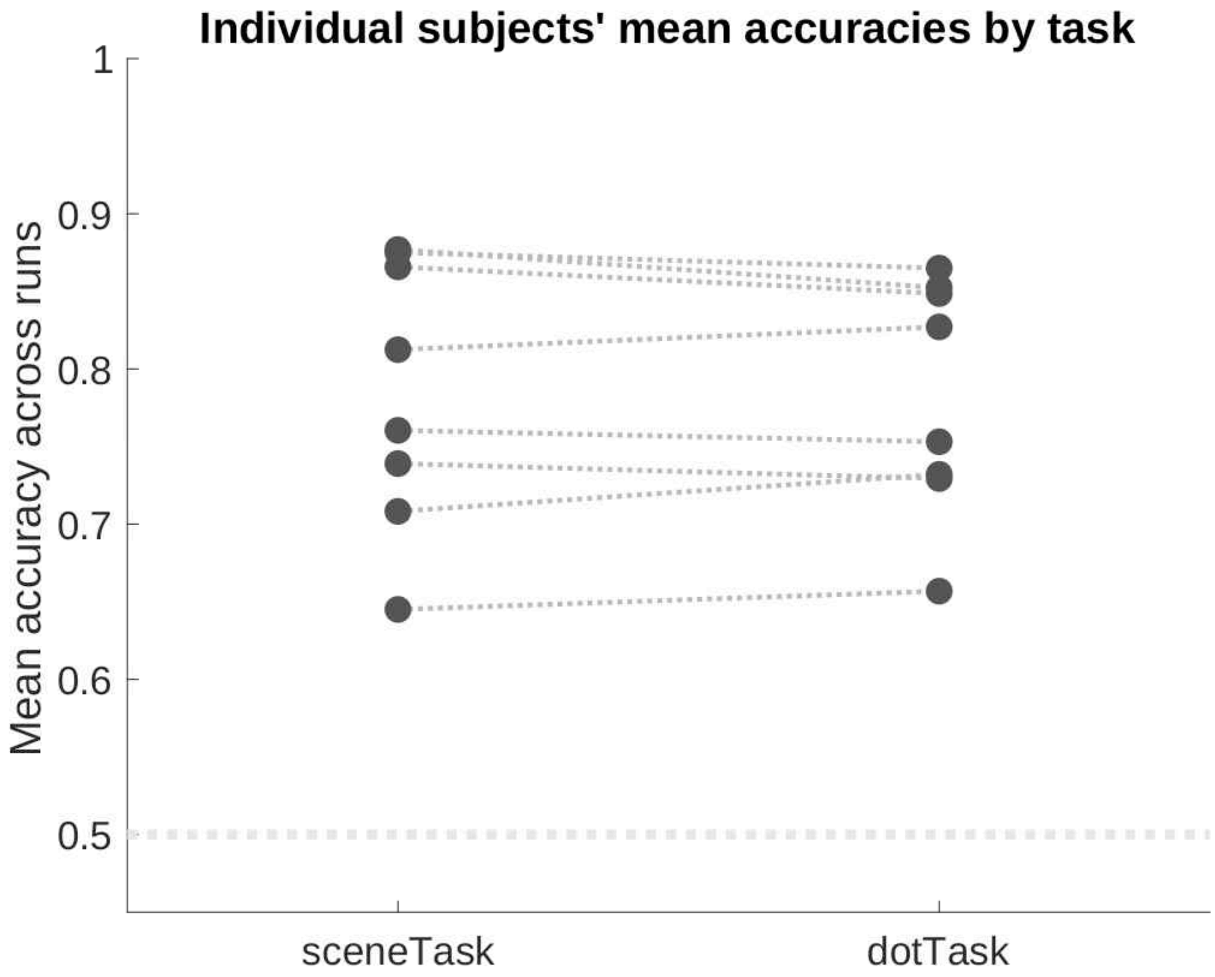


Figure S3.1. Participants' performance was well above chance, with no obvious outliers. The difficulty of the scene task was fixed because it depended on pre-determined placement of dots on the scene. We chose these placements with the goal of keeping participants off of ceiling. We were able to staircase 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 S3.1. RDM-correlations between 2D gabor-wavelet model and 3D-scene-surface (global) models are reliable within stimulus sets across 3 distance metrics. For Chapters 2 and 3, only the global scene-surface model was included, as it most closely corresponded to Lescroart & Gallant's (2019) model.

Distance metric	**Chapter 2**	**Chapter 3**	**Lescroart & Gallant (2019)**
Cosine distance	0.056	-0.002	0.274
Spearman's rho	0.052	-0.001	0.191
Pearson correlation	0.057	-0.003	0.277

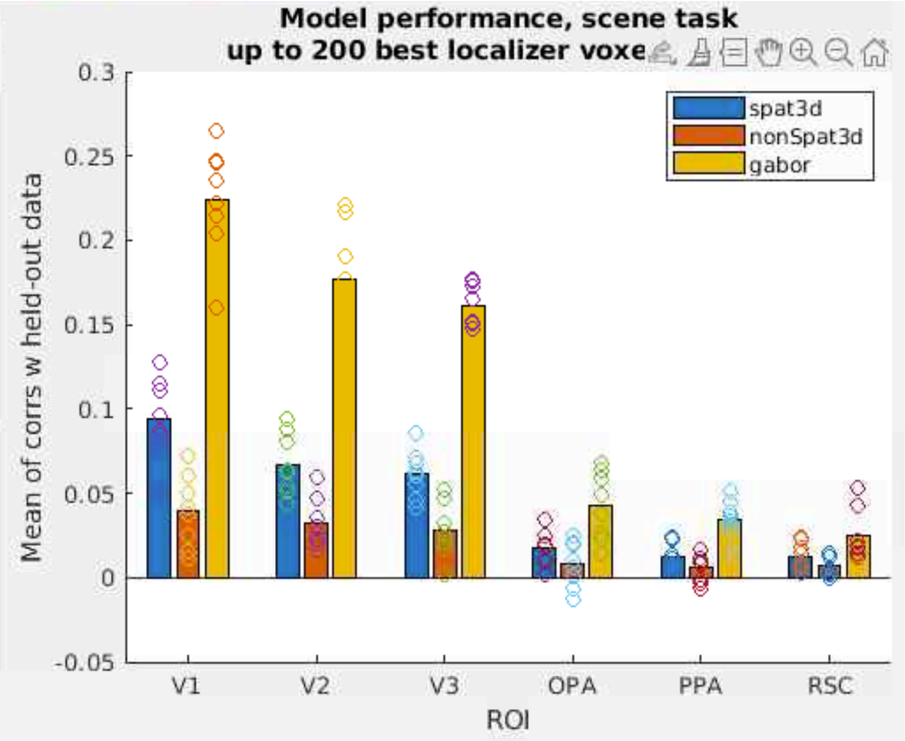


Figure S3.2: Scene-task-only performance. Magnitudes are almost identical to performance collapsed across task, suggesting that our lack of evidence for variance uniquely attributable to 3D surfaces is not due to our task manipulation.

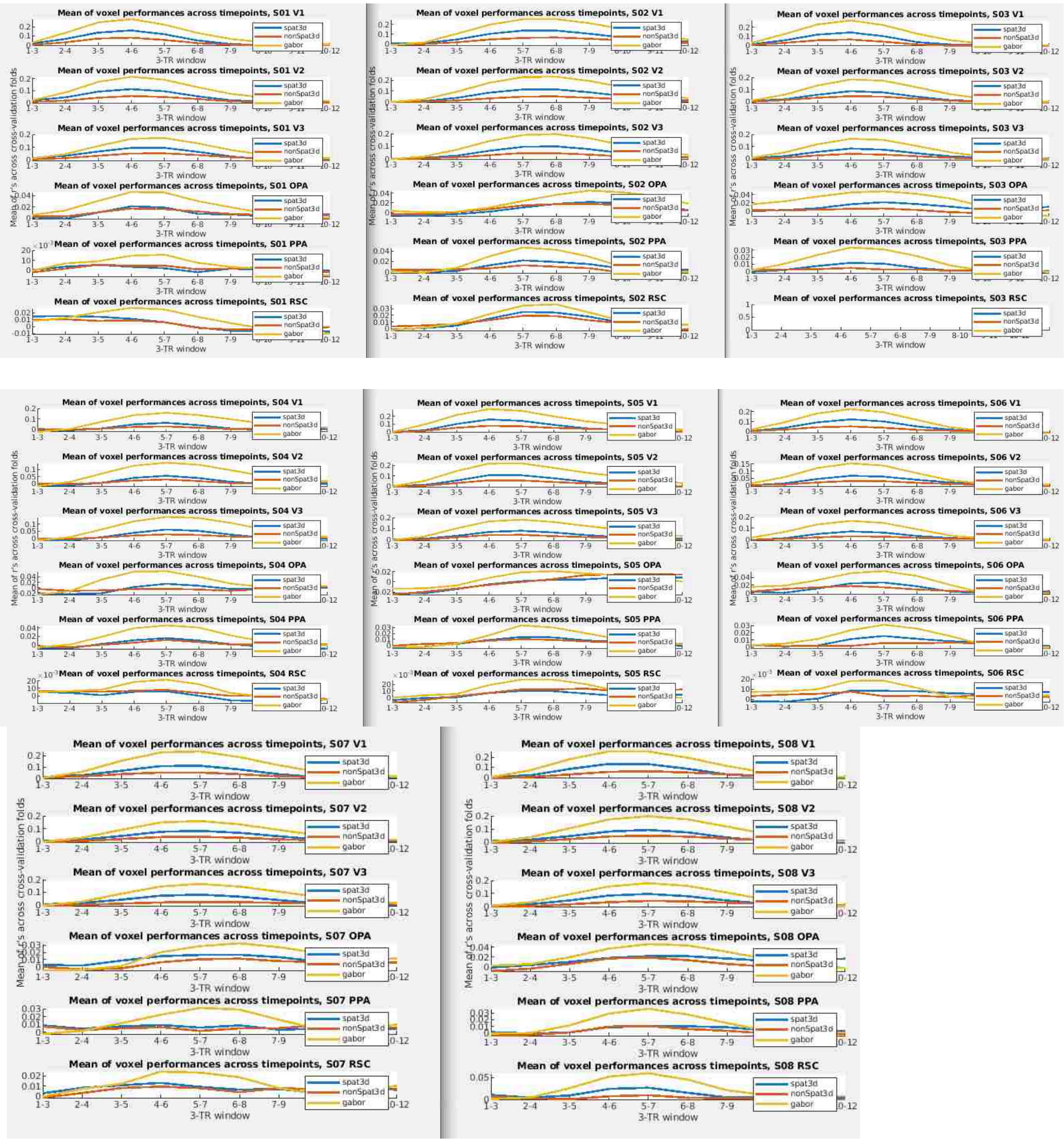


Figure S3.3 Sliding-window results. Prediction performance across ten 3-TR windows. The relative pattern of model performances is robust, even in individual subjects, arguing that our results are not due to differences in timepoints analyzed compared to other studies. The pane with no lines (S03, RSC)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

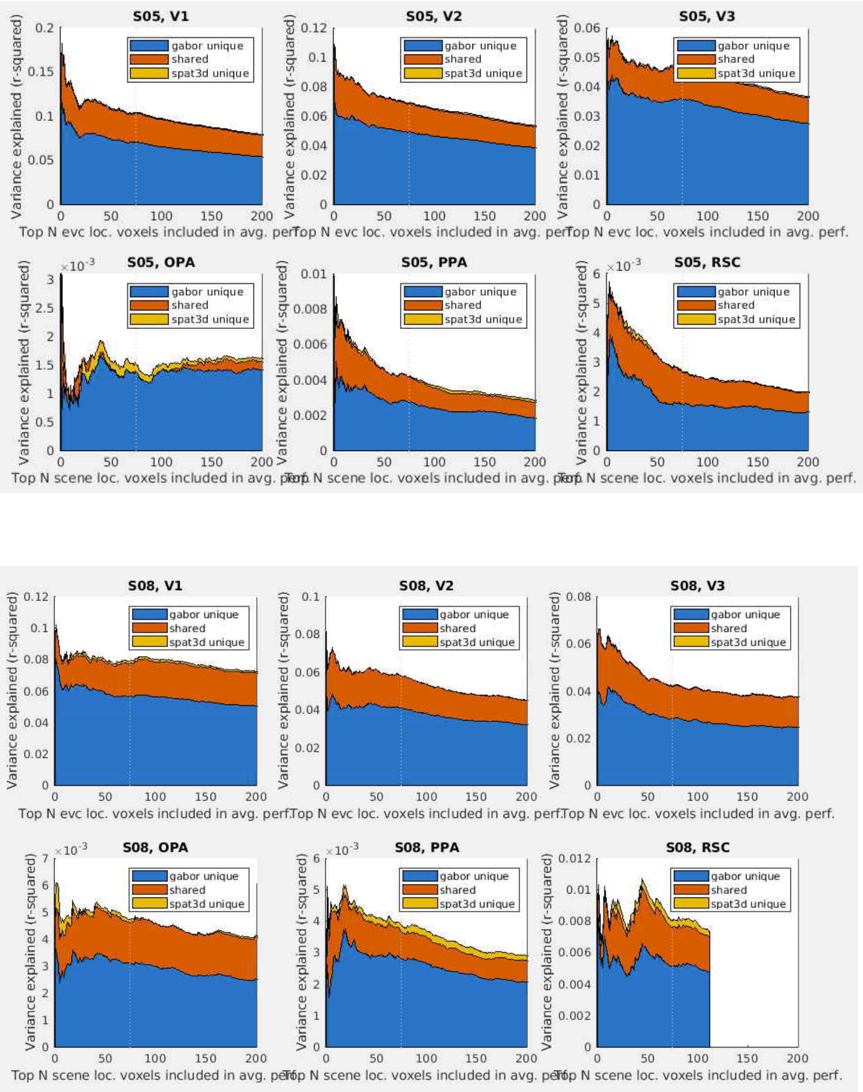
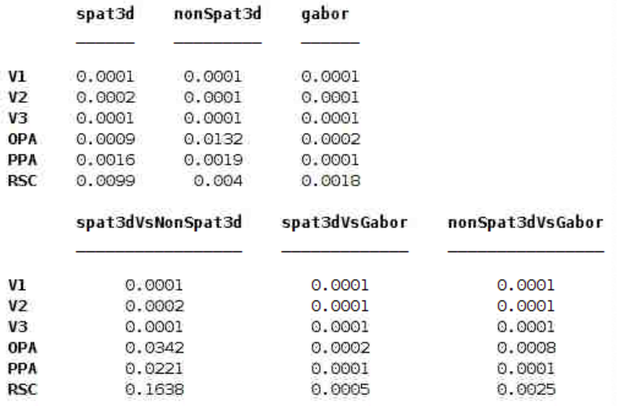


Figure S3.4: **Variance partitioning results are robust to ROI size.** Two example subjects' variance partitioning results plotted as a function of voxels included in each ROI. Voxels were ordered by independent-localizer activation: scrambled > baseline for EVC ROIs, top row; scenes > objects for scene-selective 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, y-axis-height-taken-up denoting: unique variance explained (r^2) by the gabor model (first legend entry; bottom layer of stacked plot; blue; sorry the colors don't match the main-results colors); shared variance (middle; red); and variance uniquely explained by the 3D model, in this case the spatial, or quadrant-based model (bottom legend entry; top layer of stacked plot; yellow). Note that these really _are_ representative example subjects because all subjects looked remarkably similar. 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 S3.2A: Two-tailed p-values corresponding to permutation tests for each model compared to 0, as well as comparisons of performance for each.



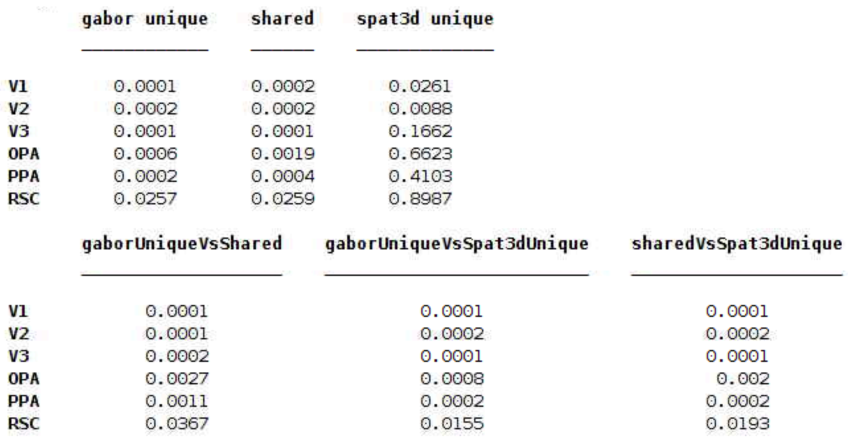
```
           spat3d     nonSpat3d     gabor
           ______     _________     ______

    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 S3.2B: Variance partitioning, two-tailed p-values corresponding to permutation tests for quadrant-based 3D ("spat3d") and gabor models.



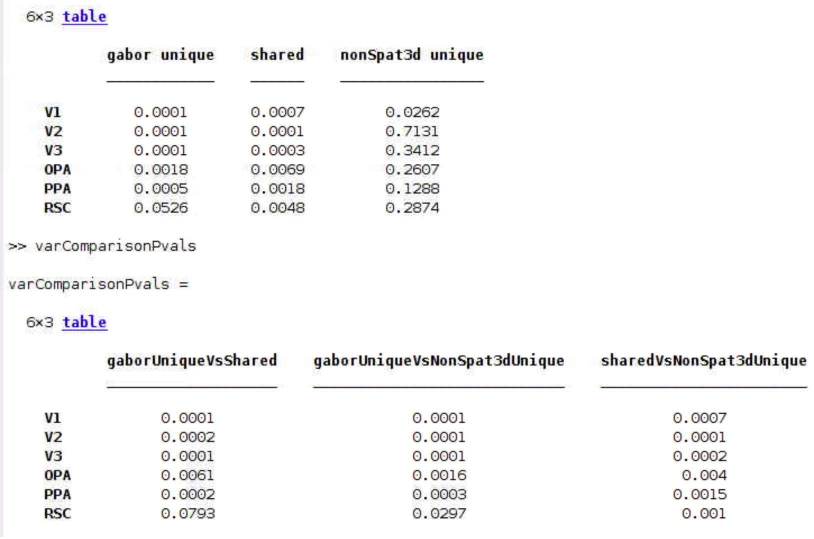
```
           gabor unique    shared    spat3d unique
           ____________    ______    _____________

    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 S3.2C Variance partitioning, two-tailed p-values corresponding to permutation tests for global 3D ("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
```