A revised framework for human scene recognition

Author: Drew Linsley

A Revised Framework for Human Scene Recognition

Drew Linsley

A dissertation

submitted to the Faculty of

the department of Psychology

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Boston College
Morrissey College of Arts and Sciences
Graduate School

May 2016

A Revised Framework for Human Scene Recognition

Drew Linsley

Advisor: Sean P. MacEvoy, Ph.D.

For humans, healthy and productive living depends on navigating through the world and behaving appropriately along the way. But in order to do this, humans must first recognize their visual surroundings. The technical difficulty of this task is hard to comprehend: the number of possible scenes that can fall on the retina approaches infinity, and yet humans often effortlessly and rapidly recognize their surroundings. Understanding how humans accomplish this task has long been a goal of psychology and neuroscience, and more recently, has proven useful in inspiring and constraining the development of new algorithms for artificial intelligence (AI). In this thesis I begin by reviewing the current state of scene recognition research, drawing upon evidence from each of these areas, and discussing an unchallenged assumption in the literature: that scene recognition emerges from *independently* processing information about scenes' local visual features (i.e. the kinds of objects they contain) and global visual features (i.e., spatial parameters). Over the course of several projects, I challenge this assumption with a new framework for scene recognition that indicates a crucial role for information sharing between these resources. Development and validation of this framework will expand our understanding of scene recognition in humans and provide new avenues for research by expanding these concepts to other domains spanning psychology, neuroscience, and AI

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DEDICATIONS

For Annie, Mom, Dad, and Jeremy.

ACKNOWLEDGEMENTS

Here is my recipe to completing a Ph.D.: mix equal parts self-sufficiency and stubbornness with a giant heaping of support from your family, friends, and colleagues. Whisk to blend. Bake for 5 years then reflect on one of the best decisions you have ever made.

Thank you to all of those who have played a part in this endeavor.

To the students, faculty, and staff of the Boston College Psychology Department, who have become my dear friends and colleagues.

To my defense committee members, Elizabeth, Liane, and Aude. Thank you for being enormously generous with your time. You are incredible role models.

To Sean, thank you for your guidance, insight, and teaching. I could not dream of a better mentor. You have shaped me into the scientist that I am today, and I am forever grateful that you took a chance on me.

To my in-laws, who have treated me like a son.

To my parents and brother, who have provided endless support, love, advice, and guidance throughout my life.

To my wife Annie. You are the love of my life, my best friend, and I could not have done any of this without you. (And Harvey, our border collie, my other best friend.)

GENERAL INTRODUCTION

Rapid and accurate recognition of the local environment is a hallmark of the human visual system and essential for a normal life. The importance of this ability is highlighted during navigation, as it allows us to identify our surroundings, determine when we have reached our destination, and engage with the local environment with appropriate behaviors along the way. A body of research amassed over the last 40 years has investigated how humans recognize scenes, and explored the types of mechanisms contributing to its effectiveness. This chapter begins with a review of this work, touching on (1) the types of scene information used during scene categorization, (2) how this information is represented in the brain, and (3) to what extent these representations ultimately influence behavior. Building upon the reviewed work, I will propose an updated framework for scene recognition that updates a long held assumption in the study of scene recognition and generates predictions for future research.

Information resources for scene categorization

Seminal studies of scene recognition have cast it as drawing on two separate sources of information. The first and most salient resource is information about the objects in a scene. Object identities are often highly diagnostic of scene identity, with many scenes closely associated with semantically related objects. For instance, stoves, ovens, and refrigerators are typically found in kitchens, whereas toilets and showers are

typically found in bathrooms. Computational research has leveraged this idea and decomposed scene recognition into a set of operations performed on scenes' objects. Through operations that extract information about the identities and spatial relationships of objects, certain algorithms are able to produce similar judgments of scene identity as humans [1–4]. Consistent with this perspective of objects as the atomic unit of scenes, human recognition accuracy is significantly reduced when a scene's objects are obscured with perceptual masks, are violating physical laws, or are in conflict with a scene's semantic context (e.g. a quarterback throwing a pass in a church; [2,5–7]). Humans are also sensitive to high-level object features, such as their physical affordances, during scene recognition [5].

The second information source for scene recognition is scenes' global or spatial properties, which describe its holistic features such as size or 3-dimensional layout [8–11]. Spatial properties are particularly useful when determining the category of a scene. For instance, an observer distinguishing between a bathroom and a kitchen can simplify this task with the insight that bathrooms are typically smaller than kitchens [12]. Although spatial properties would appear a less precise and possibly less important information source for scene recognition than objects, several pieces of evidence indicate a significant role. First, scenes can be identified approximately as quickly as the objects they contain [6,13,14], suggesting that something other than object recognition is at work. Second, observers are adept at using spatial properties to categorize scenes, with the ability to base their decisions on a variety of these features such as scenes' expanse or mean depth [10,15–17]. Third, computer vision algorithms emphasizing scene spatial properties for categorization generate similar patterns of judgments – both hits and misses

– as humans [15,18]. Despite their importance in scene categorization, a precise description of spatial properties remains nebulous, with humans showing sensitivity along multiple dimensions beyond the ones already mentioned, such as openness, navigability, and "clutter" [15,19].

Neural representations of scenes

Over the past two decades, research into human neuroscience has extensively used functional magnetic resonance imaging (fMRI). These studies often adopt a stimulus-referred approach, in which the brain is parcellated into regions with functional response profiles tuned to certain visual features more than others [20]. This approach has been particularly productive in exploring how visual information for scene recognition is represented in the brain. Most significant is the finding that the visual system contains a set of regions that are putatively sensative to either scenes' object or spatial property information – aligning with the perceptual distinctiveness of these resources and positing a neural framework for scene recognition that perhaps contains two distinct pathways.

Regions of the visual system sensitive to scene information span the occipital and temporal lobes of the brain, extending from the early visual cortex (EVC) through the ventral temporal cortex (VTC). Within VTC, lateral occipital (LO) and posterior fusiform sulcus (pF) regions of lateral occipital complex (LOC) respond significantly more strongly to images of objects than to full scenes (objects > scenes) [21–23], while the opposite pattern (scenes > objects) is observed in a region near the caudal horn of parahippocampal cortex, referred to as parahippocampal place area (PPA), as well as the retrosplenial cortex (RSC), and the transverse occipital sulcus [24–26] (TOS).

Recent methodological advances have provided a more nuanced account of the representations of object and spatial property information in these areas. This group of methods, referred to as multi voxel pattern analysis (MVPA), characterizes the spatially distributed patterns of activation in response to images of scenes (i.e. the covariance structure of neural activity). In essence, MVPA expands upon simple scalar descriptions of neural response profiles to capture their "multidimensional" sensitivity.

Using MVPA, researchers have found evidence for a complex role of LOC in scene recognition. LOC encodes representations of objects within real-world scenes that exhibit invariance to low-level visual feature transformations, such as variations in their color, shape, or luminance. Its representations also capture information corresponding to objects' semantic category, causing objects with similar affordances (e.g. a knife and cutting block) to have convergent representations [21,27–29]. Representations of objects within LOC's subfields however demonstrate different levels of invariance. Patterns of activity elicited by objects in LO but not pF show sensitivity to their position and orientation in space [30,31], and also correlate with their scene category, suggesting a potential contribution of LO to scene recognition [7].

Consistent with its robust responses to scenes, multivoxel activity patterns in PPA carry information related to both scene category and scene spatial properties. These spatial property representations span multiple spatial and textural dimensions, aligning with perceptual ratings of scenes' expanse, "spaciousness", low-dimensional estimates of their spectral information, and basic textural properties [19,32–36]. Kravitz and colleagues (2011) recently explored this sensitivity by recording activity in PPA with fMRI while participants viewed images of real-world scenes that had been ranked by

their expanse (open/closed), depth (near/far), and whether they were natural or manmade [32]. They found a striking sensitivity in PPA for scene expanse but not depth or content. Others have noted that representations in PPA are for the most part tolerant to transformations of scenes' low-level properties (e.g. line-drawings of scenes and untouched scene images are represented similarly [37]) but not to differences in scene viewing angles or content [33,38]. PPA is also sensitive to the presence and identities of objects within scenes, although the invariance of its object representations has not been extensively tested [39].

Similar to PPA, RSC patterns of activation contain information about spatial properties [39]. However, its role in the act of scene recognition remains somewhat contentious, as its representations have been considered more relevant to navigation and route-planning than scene recognition per se [40,41]. Recent work by Marchette and colleagues (2014) reinforced this distinction, showing that RSC patterns of activation maintain an allocentric, "birds eye view" map of the local environment [42]. Reinforcing this idea, RSC representations are tolerant to low- [37] and mid-level [33] transformations of scenes, as its encodings are insensitive to alterations of scene content.

Another region strongly associated with scene recognition is the TOS: it is selective to scenes' spatial properties, responding more strongly to scenes versus other types of stimuli, and has a response profile that attenuates as a function of the similarity between sequentially presented scenes, suggesting a spatiotemporal sensitivity that may be useful for recognizing scenes while navigating through the world [43–47]. Although TOS representations are invariant to low-level transformations of scene information [48,49], the characteristics of its computations have yet to be pinned down. It has been

suggested that this ambiguity is due to the proximity of TOS to multiple functionally distinct units [50] such as the inferior parietal sulcus (IPS) and superior parietal sulcus (SPS), which have recently been associated with scene categorization despite long-held associations with strictly attention-based processing [34,51].

From neural representations to behavior

That the regions discussed above encode information about scenes does not mean that they are necessary for scene recognition. For instance, the presence of this information could be epiphenomenal – feedback from some other neural process. For this reason, researchers have sought a more direct connection by measuring what happens to scene recognition performance following ablation to one of these regions (from a lesion caused by disease, stroke, or brain injury).

Patients with lesions to the lingual gyrus or parahippocampal gyrus, anatomical structures overlapping with the PPA, are often impaired in navigation and perception of environmental global properties while retaining the ability to understand spatial relationships between locations [40,52,53]. In contrast, patients with RSC lesions can still identify scenes but are unable to navigate, possibly due to a damaged allocentric perspective of their surroundings [40]. And while focal lesions to LOC are rare, case studies have demonstrated that its ablation impairs object perception while sparing egocentric spatial processing [54–58]. To my knowledge, no research exists on the effects of focal lesions to TOS on scene categorization behavior.

Transcranial magnetic stimulation (TMS) has provided an alternative method for ablating brain function by temporally "knocking out" activity in regions with extremely high spatial precision (the mechanism of function is limited to areas proximal to the

skull) [59,60]. For scene recognition research, this method has been successfully applied to infer the contributions of TOS and LO. Dilks and colleagues (2013) applied TMS to participants' TOS while they viewed images of scenes, faces, and objects [61]. They found that this manipulation caused a significant decrease in scene recognition or discriminating scene layout, but did not affect perception of other stimuli. This evidence supported the hypothesis that information processed within the TOS is causally involved with scene recognition. More recently, TMS has been used to directly link TOS to the perception of scene boundary information [62].

TMS stimulation of LO on the other hand has provided a less definitive story of its involvement in scene recognition. Mullin and colleagues (2011) reported that TMS stimulation of LO paradoxically *facillitated* reaction times during a task in which subjects indicated if scenes were manmade or natural [63]. This result contradicted fMRI evidence indicating a role (albeit an indirect one) for LO in scene categorization [7]. However, recent work indicates that Mullin's result may be a function of the type of task they used. Preliminary evidence for this is provided in a control experiment in Dilks's 2011 study, in which they saw that TMS to LO during a 4-choice scene categorization did not affect accuracy. More direct evidence has come from an investigation of the role of LO in scene recognition. We asked if LO patterns of activation could decode participants' decisions during a task in which participants categorized images of computer-generated bathrooms and kitchens, whose spatial properties made them range from easy to difficult to distinguish. We found that LO activity patterns directly impact scene category decisions *when subjects base their decisions on scenes' objects* [12]. Taken together, these results

indicate that LO's impact on scene categorization decisions may depend on the precise task asked of participants.

Neuroimaging research has identified several regions of cortex with response profiles closely tied to scene recognition. Neuropsychological research has advanced this association by demonstrated that disabling these regions produces deficits in scene recognition that align with their response profiles. PPA, RSC, and TOS encode spatial properties and impart debilitating spatial deficits when damaged, whereas scene categorization deficits are less debilitating when object-selective cortex is disabled.

Understanding biological vision through artificial intelligence

Recent years have seen an explosion of interest and applied success in the field of artificial intelligence (AI). This boom can be pinned on the extraordinary accomplishments of a simple algorithm based on neural principles, which has been fueled by advances in computational power and the availability of better data. Indeed, this field has a long history of borrowing from neuroscience [64], with algorithmic advances often coming from clever application of insights into brain function [65,66].

But the relationship between Neuroscience and AI does not run in a single direction, and has recently proven reciprocal through the application of powerful AI algorithms to help uncover the neural basis of object and scene recognition. These algorithms are broadly known as artificial neural networks (ANN), with a structure loosely inspired by observations of how biological neurons process information. Given an input (e.g. weather for the past week) and a target (e.g. what clothes were worn), a neuron in an ANN will identify a set of transformations to map the input to the target, which creates a model that can generate predictions for novel instances of the same problem.

Taking this analogy to the grand scale necessary for scene recognition, millions of artificial neurons work together to transform an image of a scene into a label such as "kitchen". Similar to the architecture of the VTC, these neurons are assembled into a hierarchical structure, in which the computation performed by each successive neuron transforms the input into a progressively more abstract representation (i.e. from information about the edges in an image to its object contents). Astoundingly, this incredibly simple abstraction of the VTC not only reaches near-human performance in many recognition tasks, but also captures representations from images that are consistent with humans. An ANN applied to images of real-world scenes learns representations that roughly align with the progression from low- to high-level features found in cortex [67]. The similarity between these representations has been quantified as a function of algorithm performance: the better the algorithm is at recognizing images, the more similar its underlying representations are to those elicited in human VTC [68–73].

In this way, the ANN offers a potentially powerful model for understanding the neural basis of scene recognition. Within this model, the response profiles of regions in the visual system (e.g. LO versus PPA) are developed through a high-level learning algorithm that adjusts them to optimize recognition performance. That every healthy human has the same set of visual regions in roughly the same areas reflects the existence of a single unique solution to this recognition problem that all brains converge upon during development (given other biological constraints such as structure and genetics).

A revised framework for scene recognition

Behavioral, neural, and computation perspectives provide a consistent perspective on scene recognition: it emerges from a feedforward sweep of computations through the

visual system, in which complementary object and spatial property resources are processed into a behaviorally useful representation of the viewed scene. But is this relatively simplistic portrayal of the visual system incomplete? Recent work examining the independence of object and spatial property information in the visual system suggests that this might be the case.

Extant theory holds that scenes' object contents and spatial properties are processed in parallel and independently in the visual system, only converging once they reach regions responsible for cognition and decision making (reviewed in [74]). However, recent evidence indicates that this dichotomy is not strictly followed by the visual system. We demonstrated that objects strongly associated with a particular scene category can bias scenes' encoded spatial properties towards values associated with that category – an effect we termed "centripetal bias" [35]. We found that images of spatially large bathrooms were perceived as significantly more "average"-sized (i.e., smaller) when their objects (toilets, showers, and sinks) were visible versus when they were obscured with wavelet-scrambled masks. Importantly, the opposite observation was also true: small bathrooms were perceived as significantly more "average"-sized (i.e., larger) when their objects were visible versus masked. This result indicates that information extracted from scenes' objects and spatial properties are first combined during perception, rather than later stages of processing when a decision of scene category is produced [75]. This finding of centripetal bias cannot be explained by existing feedforward frameworks for scene recognition.

In this thesis, I propose the alternative framework for scene recognition depicted in Figure 0.1. Consistent with a typical feedforward framework for scene recognition, the

figure shows an input image (at the top) being processed and correctly recognized as a kitchen (the label "Kitchen" at the bottom). However, this framework is drastically different than existing accounts in what happens between these two points.

The arrows extending down from the image towards the label "Kitchen" represent the information resources humans use to recognize scenes: spatial properties in red and object features in blue. Consistent with evidence of centripetal bias explained in Chapter 1, spatial property information is influenced by the identities of a scene's objects (blue "object" dots over the red spatial property). But is this influence only brought to bear from objects onto spatial properties, or is the opposite also true? In a set of studies, we demonstrate that spatial properties have the ability to "fill-in" missing object information (blue object arrow with red spatial property dots; Chapter 2) – an effect supported by a simple form of statistical learning instantiated in the visual system (grey circle; Chapter 3). Importantly, these perceptual biases are not epiphenomenal as they significantly improve scene recognition accuracy by filling in missing or obscured visual features without drawing from computationally slow cognitive feedback mechanisms (black circled X; Chapter 4). Further development and validation of this framework for scene recognition will bring us closer to understanding how humans recognize the visual world.

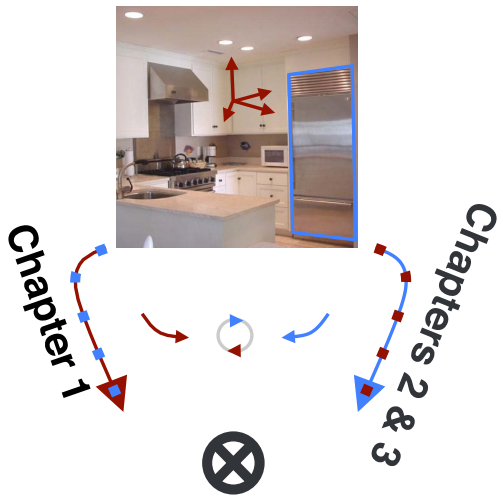


Figure 0.1. A revised framework for scene recognition. Similar to existing models, the observer recognizes the viewed scene as a kitchen through a combination of information about its objects (in blue) and spatial properties (in red). Here we provide evidence for the combination of these resources during perception, the learning mechanisms supporting this combination, and its impact on scene recognition.

1.0 ENCODING-STAGE CROSSTALK BETWEEN OBJECT- AND SPATIAL PROPERTY-BASED SCENE PROCESSING PATHWAYS

Published as: Linsley D, MacEvoy SP. Encoding-Stage Crosstalk Between Object- and Spatial Property-Based Scene Processing Pathways. Cerebral Cortex 2014.

Scene categorization is a central task of the visual system, and draws on two information sources: the identities of objects scenes contain, and scenes' intrinsic spatial properties. Because these resources are formally independent, it is possible for them to produce conflicting estimates of scene category. We tested the hypothesis that the potential for such conflicts is mitigated by a system of "crosstalk" between object- and spatial layout-processing pathways, under which the encoded spatial properties of scenes are biased by scenes' object contents. Specifically, we show that the presence of objects strongly associated with a given scene category can bias the encoded spatial properties of scenes containing them towards the average of that category, an effect which is evident both in behavioral measures of scenes' perceived spatial properties, and in scene-evoked multivoxel patterns recorded with fMRI from the parahippocampal place area (PPA), a region associated with the processing of scenes' spatial properties. These results indicate that harmonization of object- and spatial property-based estimates of scene identity begins when spatial properties are encoded, and that the PPA plays a central role in this process.

1.1 INTRODUCTION

Fast and accurate scene recognition is critical to daily life, allowing us to navigate from one place to another and interact efficiently and appropriately with the environment at each point along the way. Objects have often been cast as the fundamental building blocks of scenes, with scene recognition proposed to emerge from a cataloging of the types of objects in a scene and the spatial relationships among them [76–80]. Consistent

with this view, behavioral studies have shown that scene recognition falters when highly informative objects (e.g., refrigerators or toilets) are removed from their associated scenes [81] or when scenes contain incongruent objects [82,83]. More recently, however, theoretical and behavioral studies of scene recognition have demonstrated a substantial role for scenes' intrinsic global properties, particularly spatial properties such as depth, openness, and navigability, complementing the category cues provided by objects [84–92].

From a physical perspective, the kinds of objects scenes contain and scenes' spatial properties are often unrelated. To be sure, these factors place constraints on each other: objects help define a scene's spatial dimensions, and a scene's spatial dimensions may place limits on the types of objects it can contain. Yet many scene categories that differ from each other in their spatial properties can nevertheless accommodate each other's associated objects without grossly altering their properties or violating physical law. For instance, although bathrooms and kitchens may differ in their average dimensions at the category level, there is usually no physical (as opposed to semantic) impediment to replacing a kitchen's refrigerator with a shower, or a bathroom's bathtub with a stove, and little change in either scene's spatial properties as a consequence of the switch. In general, all other things (e.g., object size) being equal, the *identities* of objects in a scene and the scenes' spatial properties are, with a few exceptions, logically independent.

The reliance of scene recognition on two cues, objects and spatial properties, that can vary fairly freely with respect to each other raises a problem: how do scene categorization decisions cope with the potential for conflicts between the categories most

associated with each cue? Consider, for example, the problem of categorizing a large bathroom. While the objects present (a sink, toilet, and bathtub) might be closely associated with bathrooms, the spatial dimensions of the room might be more closely associated with a different room category, such as a kitchen. Drawing from models positing that both objects and scenes' spatial properties can activate schemata or context frames for specific scene categories [77,79,80,93–100], we might imagine that this conflict is resolved via some negotiation between the schemata activated by each resource. In this view, objects and spatial properties are processed independently through the stage at which each triggers a "hypothesis" of the room's category. An alternative is that potential conflicts between the room's object contents and spatial properties are blunted at the stage at which those resources are encoded; i.e., *before* schemata are activated by each. Information from whichever resource is likely to be more reliable under the circumstances (likely objects in this scenario) is used to bias the how the other resource is *encoded*, with the goal of maximizing the likelihood that each resource ultimately activates the same schema.

In the present study we used a combination of behavioral and neuroimaging techniques to assess whether any such encoding-stage bias occurs during scene viewing, specifically asking whether the encoded values of scenes' spatial properties are influenced by scenes' object contents. Taking advantage of the susceptibility of perceived spatial properties to negative aftereffects [101], we first asked participants in a series of behavioral experiments to judge the spatial scales of average-sized bathrooms and kitchens after prolonged exposure to either very large or very small scenes from the same category. We find that the magnitudes of aftereffects produced by both small and large

adapting rooms were significantly smaller when objects which were strongly informative of scene category, such as refrigerators or toilets, were visible in adapting scenes versus when they were masked. These results indicate that the presence of informative objects in adapting scenes biased their encoded spatial properties towards values associated with the "average" of their category. Next, using fMRI we observed an essentially identical bias in scene-evoked activity patterns in the parahippocampal place area (PPA), a region which has been associated with the encoding of scenes' spatial properties [102–107], as well as processing of scenes' contextual associations [108–111]. We propose that these behavioral and physiological biases reflect "crosstalk" between object- and spatial property-processing pathways that serves scene recognition by harmonizing estimates of scene identity derived from each of these information resources, and that this process is mediated by PPA.

1.2 MATERIALS AND METHODS

1.2.1 Behavioral Experiments

1.2.1.1 Participants

A total of 119 participants (25 male) between 18-27 years old were recruited for the study, chiefly from among Boston College undergraduates enrolled in introductory psychology courses. All had normal or corrected-to-normal vision and provided written informed consent in accordance with the procedures of the Boston College Institutional Review Board. Participants were either paid $15 or received course credit.

1.2.1.2 Stimuli

We assessed the influence of objects on scenes' encoded spatial properties by measuring changes in the magnitudes of negative aftereffect produced by adaptation to scenes with extreme spatial properties varied with the visibility of objects within them. Visual stimuli were 500 full-color photographs of real-world bathrooms and kitchens (1,000 photographs total), and 500 computer-rendered images of exemplars of kitchens (Figure 1.1.A; see Appendix A Figure 1 for additional exemplars). Bathrooms and kitchens were selected for use because they are strongly associated with distinct sets of objects, as judged by the authors, and because, as indoor scenes, they have clearly identifiable bounding features (such as walls) that make relative judgments of spatial scale intuitive to naïve observers. Real-world scenes were assigned to quintiles according to perceived size (hereafter referred to as "spaciousness"), as judged by web-based ratings made by a pool of paid raters recruited through Amazon Mechanical Turk (AMT), while rendered scenes were assigned to size quintiles according to their simulated floor areas. Scenes from the first and fifth quintiles (i.e., those that were very small and very large for their category, respectively) were selected for use as spatially-extreme adapting scenes, and scenes in the middle three quintiles were designated for use as spatially-average "test" scenes against which the effects of adaptation would be measured. Please see Appendix A for details of the AMT scene rating procedure and methods of rendered scene generation.

Because we wished to understand how the objects in scenes might bias scenes' encoded spatial properties, we generated a second version of each adapting scene in which informative objects were obscured. Informative objects were identified based on

nominations made by each AMT rater of the three objects he or she most strongly associated with bathrooms and kitchens, with the three objects nominated objects most frequently for each category selected for use. Informative objects for bathrooms were toilets, sinks, and bath/showers; for kitchens they were refrigerators, sinks, and stoves/ovens. Within each real-world scene, the spatial boundaries of as many as visible of the three appropriate objects were segmented with the LabelMe image annotation tool [112], and a "masked" version of each scene was generated by replacing those objects with scrambled biorthogonal 3.1 discrete wavelet transforms [113]. On average, the number of informative objects masked in each scene was 2.89 for high-spaciousness bathrooms, 2.31 for low-spaciousness

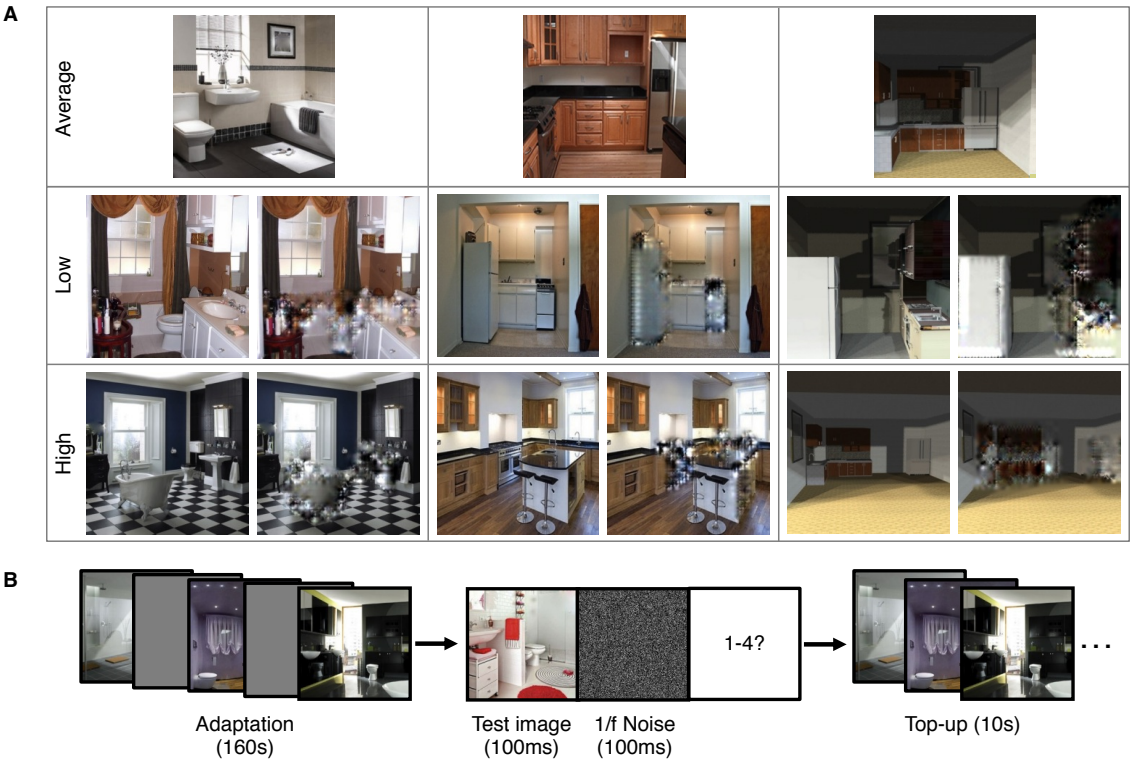


Figure 1.1. Experimental stimuli and design (A) Participants judged the spaciousness of images of exemplars of (top row, left to right) average-sized bathrooms, kitchens and computer rendered kitchens. Judgments were preceded by adaptation to low- or high-spaciousness exemplars from the same categories (second and third rows, respectively) that had informative objects either unmasked or masked (left and right images in each cell, respectively). (B) Each adaptation block began with 800 adapting scenes (100 exemplars repeated 8 times, totaling 160 seconds) during which participants performed a one-back repetition detection task to ensure attention. Participants then rated the spaciousness of 30 test scenes. Test scenes always had all informative objects unmasked, and were preceded by a 10 second top-up adaptation period containing 100 images from the initial adaptation period.

bathrooms, 2.98 for high-spaciousness kitchens, and 2.54 for low-spaciousness kitchens. Objects in rendered scenes were masked in the same way.

1.2.1.3 Procedure

Behavioral adaptation experiments were created and executed in MATLAB using the Psychophysics Toolbox [114]. All scenes subtended 10° square when viewed from a chinrest positioned 57cm from an LCD monitor. Each participant judged the spaciousness of test scenes in a series of adaptation blocks. Following the design of a previous study demonstrating negative aftereffects in judgments of scenes' global properties after adaptation to scenes with extreme properties [101] (Figure 1.1.B), each block began with 800 exemplars from either the high- or low-spaciousness quintiles of a single scene category, corresponding to 8 presentations of each of the 100 exemplars from the appropriate quintile, in random order. Each adapting scene was shown for 100ms and followed by a 100ms gray screen, with exact timing yoked to the 60 Hz frame rate of the monitor. During adaptation sequences, participants were asked to perform a 1-back repetition detection task to maintain attention. Following the initial adaptation period, participants reported the subjective spaciousness of test scenes from either the same or different category. Each test scene was shown for 100ms followed by a 100ms 1/f noise mask, after which participants rated the scene's spaciousness according to the scheme outlined in the following paragraph. To maintain a consistent adaptation state, test scenes were separated by "top-up" adaptation periods consisting of 100 adapting scenes shown for 100ms each without interruption over 10s. Each participant rated 30 test scenes in each of four adaptation blocks corresponding to each combination of adapting scene spaciousness (high versus low) and informative object masking state (all unmasked

versus all masked). Each block used a unique set of randomly selected test scenes, equated for average spaciousness across blocks. Each test scene was viewed by a participant only once.

In the first experiment undertaken, featuring adaptation to high- and low-spaciousness bathrooms, participants rated the spaciousness of test bathrooms on a 5-point scale (1 = "much less spacious than average ", 2 = "less spacious than average", 3 = "about average spaciousness", 4 = "more spacious than average", 5 = "much more spacious than average"); ratings in all subsequent experiments were on a 4-point scale that was identical but for elimination of the middle "about average" option. This change was an attempt to improve the sensitivity of the rating scale to small changes in perceived spaciousness. Specifically, we were concerned that the "about average" response option in the 5-point scale was conceptually too broad, and therefore might be chosen by participants even for scenes which they felt were actually slightly above or below average in spaciousness. The 4-point scale theoretically avoided this problem by forcing participants to make a "more" versus "less" decision for every scene. Because the aim of the 4-point scale was to improve sensitivity to aftereffects that were already statistically significant in the initial bathroom experiment using the 5-point scale (see Results), repetition of that experiment with 4-point scale was not necessary.

For reasons outlined in the Results, adaptation experiments were executed with five non-overlapping groups of participants. For the first three groups, the design was exactly as described above: participants rendered spaciousness judgments of scenes from a single category after adaptation to scenes from the same category. The categories used in these three experiments were bathrooms, kitchens, and computer-rendered kitchens;

participants in each of these experiments are referred to hereafter as the "real bathroom" (participant $n = 35$), "real kitchen" ($n = 17$) and "rendered kitchen" ($n = 30$) groups, respectively. As is described in the Results, the small size of the real kitchen group reflects early termination of that experiment when it became clear that the results were inconsistent with our hypothesis due to potential confounding factors in those photographs.

To understand whether adaptation effects were category specific, two additional groups of participants judged scenes from one category of real-world scenes (kitchens or bathrooms) after adaptation to scenes from the *other* category. The first of these groups were subjected to a variant of the main experiment design in which they rated kitchens after adaptation to bathrooms, and vice versa, but with objects always unmasked in all scenes. This experiment was designed to measure whether "basic" aftereffects (i.e., different test scene judgments after adaptation to high- versus low-spaciousness scenes) could be observed between categories; participants in this experiment will be referred to as the "unmasked cross-category" group ($n = 18$). The small size of this participant group reflects the generally large magnitude of basic adaptation effects. Participants in the second "cross-category" group rated kitchens after adaptation to bathrooms that either had objects masked or unmasked. Participants in this experiment will be referred to as the "masked cross-category" group ($n = 29$).

1.2.1.4 Data Analyses

Due to the evanescence of aftereffects [115], we wished to exclude delayed spaciousness ratings. To do so, and to exclude trials in which participants responded implausibly

quickly, data for each participant were scanned for reaction times (RTs) and trials were excluded from analysis when RTs fell more than 4 standard deviations above or below the mean of trials accumulated across all members of his or her participant group. We also employed a clustering algorithm to provide an unbiased means of filtering responses of inattentive participants. Details of both of these procedures can be found in the Appendix A Table 1. Note that these procedures reduced the number of participants in each group who contributed data to final analyses. Final participant counts can be found in the Results.

1.2.1.5 Statistical Analyses

We used a series of permutation tests to assess the statistical significance of differences in judgments of test scene spaciousness among adapting scene types. For each comparison of interest between two types of adapting scenes (e.g., spaciousness ratings after adaptation to low- versus high-spaciousness scenes), we randomly permuted the condition labels of the 60 spaciousness judgments each participant made across the two types of adapting scenes. The difference between the average ratings in each pair was recomputed based on the permuted labels, and the average value of this difference across participants was stored. This permutation procedure was repeated 10,000 times, allowing us to construct the distribution of group-level rating differences to be expected under the null hypothesis that test scene spaciousness ratings were unaffected by adapting scenes. The *p* value of the actual rating difference between the adaptation conditions being compared was the proportion of elements in the null distribution that exceeded the actual rating difference. Permutation testing was selected to avoid distributional assumptions of parametric tests.

Separate planned tests were performed to assess the significance of differences in aftereffects within three pairings of adapting scene types. Each test involved its own independent set of label permutations. In the first test, labels were permuted between ratings following adaptation to high- and low-spaciousness unmasked adapting scenes in order to measure the strength of "basic" aftereffects that had been reported previously for spatial properties [116]. In the second test, labels were permuted between ratings following adaptation to high-spaciousness masked and unmasked adapting scenes in order to measure the effect of object masking on aftereffects produced by high-spaciousness scenes. Finally, in the third test labels were permuted between ratings following adaptation to low-spaciousness masked and unmasked adapting scenes to measure the effect of object masking on aftereffects produced by low-spaciousness scenes. Because we had a clear hypothesis about the sign of the rating difference for each of these comparisons, *p* values were determined from one tail of null permutation distributions.

1.2.2 fMRI EXPERIMENT

1.2.2.1 Participants

Thirteen participants (12 female, aged 18-23 years) with normal or corrected-to-normal visual acuity gave written informed consent in compliance with procedures approved by the Boston College Institutional Review Board. One participant with excessive motion artifacts was excluded from analysis. Participants were paid $45.

1.2.2.2 Stimuli

Visual stimuli were real-world bathroom image exemplars assembled for the behavioral experiment, except in gray- or blue-scale format (Appendix A Figure 3). Both masked and unmasked versions of these images were used. Scenes subtended 9.3° of visual angle.

1.2.2.3 Experimental Procedure

Each stimulus event consisted of a bathroom image presented for 150ms, followed by a white fixation cross for 1350ms. Five types of bathrooms were shown: exemplars from the top and bottom spaciousness quintiles shown both with informative objects unmasked and masked, and exemplars from the middle quintile with objects unmasked. Participants indicated the color of the bathroom (gray or blue) by button press when the fixation cross appeared. The five scene types along with 3-second null events were ordered according to third-order counterbalanced de Bruijn sequences, a general class of pseudorandom sequences which provide the minimum length sequence needed to achieve a desired depth of stimulus counterbalance for a condition set of arbitrary size [117,118]. Each scan run contained 36 repetitions of each stimulus type. Runs lasted 6 minutes and 18 seconds, including 15-second fixation-only intervals attached to the end of each run. Unique stimulus sequences were constructed for six scan runs for each subject. Scan sessions also included two functional localizer scans lasting 7 minutes 48 seconds each, during which subjects viewed blocks of color photographs of scenes, faces, common objects and scrambled objects presented at a rate of 1.33 pictures per second [119]. Localizer stimuli subtended 15° of visual angle.

1.2.2.4 MRI Acquisition

All scan sessions were conducted at the Brown University MRI Research Facility using a 3T Siemens Trio scanner and a 32-channel head coil. Structural T1* weighted images for anatomical localization were acquired using a 3D MPRAGE pulse sequences (TR = 1620 ms, TE = 3 ms, TI = 950 ms, voxel size = 0.9766 x 0.9766 x 1mm, matrix size = 192 x 256 x 160). T2* weighted scans sensitive to blood oxygenation level-dependent (BOLD) contrasts were acquired using a gradient-echo echo-planar pulse sequence (TR = 3000ms, TE = 30ms, voxel size = 3x3x3mm, matrix size = 64 x 64 x 45). Visual stimuli were rear projected onto a screen at the head end of the scanner bore and viewed through a mirror affixed to the head coil. The entire projected field subtended 24° x 18° at 1024 x 768 pixel resolution.

1.2.2.5 fMRI Analysis

Functional images were corrected for differences in slice timing by resampling slices in time to match the first slice of each volume, realigned with respect to the first image of the scan, and spatially normalized to the Montreal Neurological Institute (MNI) template. Volumes from experimental scans were analyzed with general linear models (one for each scan run) implemented in SPM8 (http://www.fil.ion.ucl.ac.uk/spm), including an empirically-derived 1/f noise model, filters that removed high and low temporal frequencies, and nuisance regressors to account for global signal variations, between-scan signal differences, and participant movements. Beta value maps were extracted for each stimulus condition for each scan.

Next, within each subject a permutation test was used to identify voxels whose responses varied significantly among scene types and thus would be passed to the multivoxel pattern analysis (MVPA) used for hypothesis testing. For each voxel we stored the F statistic from a one-way ANOVA performed on beta values from each of the five bathroom types, sampled across the six scans. This statistic was compared to a null distribution of F statistics computed from 10,000 within-scan permutations of condition labels, accumulated across all voxels. A voxel was passed to subsequent analysis if its unpermuted F statistic exceeded the 95th percentile of the null distribution. Selection based on a null distribution accumulated across all voxels was a conservative approach, ensuring that only the voxels with responses differing most consistently across conditions were selected for further analysis. Note, however, that while this procedure identified voxels whose responses varied among stimuli, it was not biased towards identifying voxels with *any particular ordinal relationships* among those responses. That is, a voxel could satisfy our selection criterion as easily with responses to stimuli labeled A, B,C,D and E that reliably fell in the order A>B>C>D>E as with responses that were reliably ordered B>C>A>E>D, or any other order. Because our hypotheses addressed ordinal relationships, as explained in the following paragraph, our ANOVA-based feature selection procedure thus did not amount to "peeking". We required that at least 7 voxels be selected from each region of interest (ROI; see below for definitions) in each participant. This minimum was selected because it equaled the number of voxels in each searchlight cluster used for whole-brain analyses, as described in a following section. Response vectors composed of selected voxels were generated for each stimulus type and

averaged across scans, and pairwise Euclidean distances among all vectors were computed for each participant.

1.2.2.6 Statistical Analysis

To test the hypothesis that informative objects bias patterns of neural activity evoked by high- and low-spaciousness bathrooms, we used permutation tests to assess the group-level significance of series of contrasts among pairwise Euclidean pattern distances for each ROI. First, we needed to ask whether patterns in each ROI were sensitive to scenes' actual properties, without reference to any potential effects of object. An ROI was considered sensitive to spatial properties if it showed a significantly positive value for the distance contrast [(*distance from unmasked high-spaciousness scenes to unmasked low-spaciousness scenes*) minus (*average of distance from unmasked high-spaciousness to average-spaciousness and from unmasked low-spaciousness to average-spaciousness*)]. The significance of this contrast was measured via 10,000 within-scan permutations of the condition labels for unmasked high-spaciousness, unmasked low-spaciousness, and average-spaciousness scenes, with the contrast computed for each permutation to generate a distribution of values expected under the null hypothesis that patterns were not related to scenes' spatial properties. Second, to measure any biasing effect of objects on patterns evoked by low-spaciousness scenes relative to average-spaciousness scenes we computed the contrast [(*distance from masked low-spaciousness to average-spaciousness*) minus (*distance from unmasked low-spaciousness to average-spaciousness*)]. The significance of this contrast was tested by permuting labels for unmasked low-spaciousness and masked low-spaciousness scenes. Third, to measure any biasing effect of objects on patterns evoked by high-spaciousness relative to average-

spaciousness scenes we computed the contrast [(*distance from masked high-spaciousness to average-spaciousness*) minus (*distance from unmasked high-spaciousness to average spaciousness*)]. The significance of this contrast was tested by permuting labels for unmasked high-spaciousness and masked high-spaciousness scenes. Finally, to provide context for any of significant values for the above contrast, we computed the contrast [(*distance from masked high-spaciousness to masked low-spaciousness*) minus (*distance from unmasked high-spaciousness to unmasked low-spaciousness*)]. The significance of this contrast was tested by simultaneously permuting labels between masked and unmasked high-spaciousness scenes and between masked and unmasked low-spaciousness scenes. The utility of each of these contrasts is explained further in the Results. Because we had clear hypotheses about the sign of each contrast, one-tailed tests were applied, with values from unpermuted data considered significant if they were exceeded by fewer than 5% of permuted values. Our focus on specific distance contrasts obviated the need for cross-validation that is often employed with MVPA [120,121].

Whole-brain searchlight pattern analysis was performed with 3 mm radius (7 voxel) searchlights centered on each voxel in the brain [122]. To measure any local biasing effect of object visibility on scenes' encoded spatial properties, Euclidean distances among patterns evoked by each stimulus at each searchlight position were used to compute the distance contrast [(*distance from masked high-spaciousness to masked low-spaciousness*) minus (*distance from unmasked high-spaciousness to unmasked low-spaciousness*)]. The resulting value for each searchlight cluster was assigned to the voxel at its center. Resulting single-participant contrast volumes were passed to a second-level exact permutation test implemented with SnPM (http://go.warwick.ac.uk/tenichols/snpm)

and custom MATLAB scripts to assess the group-level significance of regions showing large subject-averaged contrast values, which were consistent with a biasing effect of objects. First, voxel-wise variance was smoothed with a 3 mm FWHM Gaussian filter under the nonparametric assumption of smooth underlying variance in the searchlight volumes [123]. Smoothed variance maps were used to compute maps of pseudo *t* values for each of the 2^{12} sign permutations of the 12 single-subject contrast volumes, as well as for the original, unpermuted set of volumes. The resulting 4096-element distributions of pseudo *t* values for each voxel were used to identify voxels in each permuted volume whose pseudo *t* values were encountered with a probability less than 0.001, and the size of the largest six-connected cluster of such voxels recorded for each permutation volume. Clusters identified in the same way from unpermuted volumes were considered significant if their sizes were exceeded by fewer than 5% of elements in the distribution of maximum sizes accumulated across permuted volumes. The thresholded second-level volume was projected onto a surface based representation of the MNI canonical brain with the SPM Surfrend toolbox (http://spmsurfrend.sourceforge.net), and then rendered in NeuroLens (http://www.neurolens.org).

1.2.2.7 Regions of Interest

All ROIs were defined according to a recently-described algorithmic approach applied to data from localizer scans [124]. Briefly, for each contrast of interest (e.g., scenes > than objects), a whole-brain group volume was created in which each voxel was tagged with the proportion of subjects in which that voxel showed activation exceeding a threshold of $t = 1.6$. A 3mm FWHM Gaussian filter was applied to this volume, followed by a watershed algorithm with an 8-connected parts filter applied to each axial slice. The

resulting volumes contained segmented parcels corresponding to activations shared between subjects. To reduce extraneous activations present in the segmented group volumes, parcels generated from the activations of fewer than 50% of subjects were removed. Individual subject ROIs associated with a given contrast were defined from the intersection between the shared activation volume and each subject's contrast map thresholded at $t = 1.6$. This procedure was performed for the contrasts of scenes > objects (to identify PPA, retrospenial complex (RSC), and transverse occipital sulcus (TOS)), objects > scrambled objects (lateral occipital (LO) and posterior fusiform sulcus (pFs) subdivisions of the lateral occipital complex (LOC)), and scrambled objects > objects (early visual cortex (EVC)). All voxels identified by the scenes > objects contrast that were inferior to the splenium of the corpus callosum were assigned to PPA, and all superior voxels assigned to RSC. An 11-voxel region of overlap between the group-defined candidate regions for right PPA and right pFs was assigned to PPA.

1.3 RESULTS

1.3.1 Behavioral Experiments

To measure the impact of informative objects on scenes' encoded spatial properties, participants were asked to rate the subjective spaciousness of "test" exemplars of bathrooms and kitchens that possessed spatial properties at or near the average for their category, after adaptation to exemplars which were much more spacious or much less spacious than their category average (see Methods). In the critical experimental

manipulation, adapting scenes either had informative objects unmasked (i.e., fully visible) or masked. We reasoned that any effect of informative objects on adapting scenes' encoded spatial properties should have been evident as a difference in the magnitude of aftereffects observed when objects were masked versus unmasked. This adaptation-based approach was selected over the alternative of having participants directly rate the spatial properties of scenes with and without masked objects because it avoided potential variability in participants' interpretation of masks when rating scenes with masked objects. For instance, some participants could have interpreted masks as unspecified objects, while others might have interpreted them as empty space. Although this potential problem might have been addressed via appropriate instructions, it was more cleanly avoided using an adaptation approach in which participants were never forced to make judgments of manipulated scenes. The perceptual quantity of "spaciousness" was selected as the dependent measure because it is an easily understood concept that captures the scenes' general spatial scales; we do not assert that spaciousness is a fundamental dimension along which scenes are encoded by the visual system, and acknowledge that it likely draws upon a number of more basic spatial properties which have been characterized previously [86–89].

Consistent with the susceptibility of scene spatial properties to adaptation [116], participants in the real bathroom group (n=34) rated average bathrooms to be significantly less spacious after adaptation to high-spaciousness bathrooms than after adaptation to low-spaciousness bathrooms, all with objects unmasked (Figure 2A, vertical difference between data points on the left; one tailed permutation test, $p = 0.0001$). The presence of this "basic" aftereffect demonstrates 1) that the perceptual quantity of

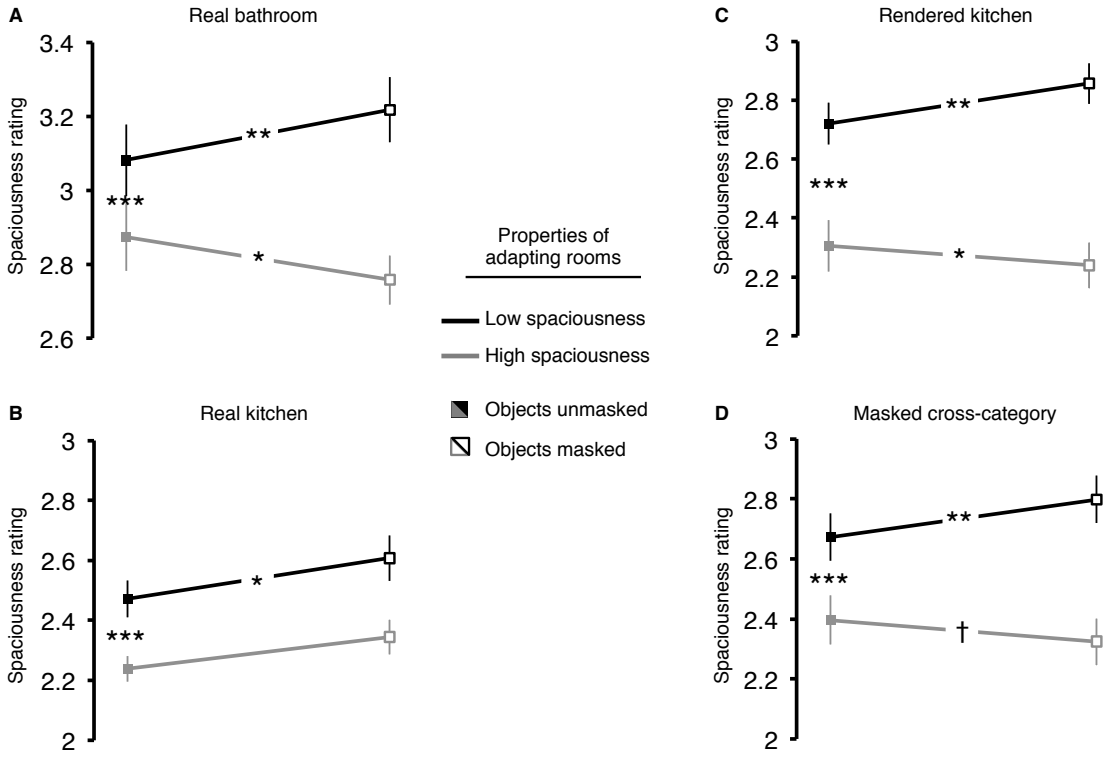


Figure 1.2. Behavioral results. (A) Average-spaciousness test bathrooms were judged significantly smaller after adaptation to high-spaciousness bathrooms than after adaptation to low-spaciousness bathrooms (black versus gray filled squares). Spaciousness ratings after adaptation to high-spaciousness bathrooms were significantly lower when adapting scenes' informative objects were masked. Object masking in low-spaciousness adapting scenes produced the opposite effect. (B) Real kitchens were similarly subject to basic (i.e., high versus low) aftereffects, although a significant impact of informative objects was only present in aftereffects produced by low-spaciousness exemplars. (C) Bidirectional enhancement of aftereffects was restored in computer-rendered kitchens, which allowed exact specification of object contents. (D) Bidirectional enhancement was also evident in ratings of kitchens after adaptation to extreme bathrooms. Error bars are s.e.m.; †, $p = 0.065$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

"spaciousness" is subject to aftereffects similar to those described previously for individuated spatial properties of scenes, and 2) that aftereffects can be observed even within the spatial constraints of a single indoor scene category.

When informative objects in high-pole adapting bathrooms were masked, aftereffects were significantly enhanced: spaciousness ratings of average bathrooms were significantly lower after adaptation to high-spaciousness bathrooms with masked objects than after adaptation to the same scenes with unmasked objects (Figure 1.2.A, vertical difference between points connected by lower line; $p = 0.018$). Based on the general observation that negative aftereffects for high-level visual features generally increase with the perceptual distance between adapting and test stimuli [125–128], this indicates that large adapting bathrooms were encoded as more spacious when informative objects were masked versus when they were unmasked. Critically, this increase in encoded spaciousness did not simply reflect space "freed up" by object removal, as evident in the fact that the magnitude of the aftereffect produced by low-spaciousness bathrooms was also significantly enhanced by object masking (Figure 1.2.A, vertical difference between points connected by upper line; $p = 0.002$), and that this enhancement took the *opposite sign*. In sum, these results indicate that *large* bathrooms were encoded as *smaller* and *small* bathrooms encoded as *larger* when informative objects were unmasked versus masked. In other words, adapting bathrooms at both spatial extremes were encoded as more similar to the average bathroom when their informative objects were visible.

Data from the real kitchen participant group ($n = 16$) showed the presence of a basic adaptation effect for kitchens (ratings after adaptation to low-spaciousness unmasked adapting scenes > after adaptation to high-spaciousness unmasked adapting

scenes = 0.0002). Furthermore, as with bathrooms, test kitchen spaciousness ratings after adaptation to low-spaciousness exemplars were significantly higher when adapting kitchens had objects masked than when they were unmasked ($p = 0.013$), indicating that object visibility biased adapting scenes to be encoded as more spacious (i.e., more similar to average kitchens; Figure 1.2.B). Unlike adaptation with bathrooms, however, aftereffects produced by high-spaciousness kitchens were not enhanced when objects were masked. (The relatively small size of this participant group was a result of our decision to terminate data collection after it became clear that there was no trend whatsoever towards enhanced aftereffects with object masking in high-spaciousness adapting scenes.)

One explanation for the absence of aftereffect enhancement with high-spaciousness kitchens is that the extra space in large kitchens allowed them to accommodate a greater number of objects carrying information about scene category than low-spaciousness kitchens, potentially blunting the impact of masking the fixed set of informative objects we targeted for masking; this potential complication applied less to bathrooms because they were associated with fewer informative objects to begin with (Table 1.1). Consistent with this explanation, high-spaciousness kitchens contained significantly more of the objects in Table 1 than did high-spaciousness bathrooms (6.99 versus 5.01 objects per scene on average; $t(198) = 11.84$, $p < 0.0001$). Moreover, we observed that large kitchens often contained many objects that, while not appearing on the list in Table 1, may still have been associated with kitchens (e.g., kitchen counter stools). Large bathrooms did not appear to collect extra potentially informative objects in a similar way.

Table 1.1. List of all objects nominated by online raters as associated with bathrooms and kitchens

Bathrooms	**Kitchens**
Bathtub	Cabinet
Mirror	Countertop
Shower	Dish
Sink	Dishwasher
Toilet	Microwave
Towel	Oven
Vanity	Refrigerator
	Sink
	Stove
	Table
	Utensil

To avoid this potential confound, we repeated the kitchen adaptation experiment with a group of participants (n = 25) who viewed computer-rendered kitchens in which spatial parameters and object contents could be exactly and independently specified. As with real kitchens, average-sized rendered kitchens were susceptible to basic aftereffects ($p = 0.001$). Moreover, aftereffects were bidirectionally enhanced when objects were masked (Figure 1.2.C; ratings after adaptation to masked high-spaciousness scenes < after adaptation to unmasked high-spaciousness scenes, $p = 0.025$; ratings after adaptation to masked low-spaciousness scenes > after adaptation to unmasked low-spaciousness scenes, $p = 0.002$), indicating that both low- and high-spaciousness adapting kitchens were encoded as more similar to average kitchens when informative objects were unmasked.

To understand whether aftereffects required that adapting and test scenes belong to the same category, participants in the unmasked cross-category group (n = 17) rated

the spaciousness of scenes from one category after adaptation to extreme unmasked exemplars from the other. Adaptation to real-world bathrooms induced significant aftereffects in real-world kitchens and vice versa (kitchen ratings after adaptation to high-spaciousness bathrooms < after adaptation to low-spaciousness bathrooms, $p = 0.0001$; bathroom ratings after adaptation to high-spaciousness kitchens < after adaptation to low-spaciousness kitchens, $p = 0.012$). These cross-category aftereffects were enhanced by object masking (Figure 1.2.D), as demonstrated in the masked cross-category participant group ($n = 24$), who showed significantly greater aftereffects in ratings of real average-spaciousness kitchens after adaptation to low-spaciousness bathrooms with objects masked ($p = 0.003$), and marginally greater aftereffects after adaptation to high-spaciousness bathrooms with objects masked ($p = 0.065$). Because of the absence of an effect of object masking on aftereffects in the real kitchen participant group, we did not examine the impact of object masking state on aftereffects produced by real-world kitchen adapters on bathrooms.

1.3.2 fMRI Experiment

Our behavioral results indicate that the presence of objects strongly associated with a particular scene category produces a "centripetal bias" in the encoded spatial properties of scenes containing them. To understand where in the visual system this bias arises, we used fMRI to record activity patterns evoked by exemplars of each of the five types of bathrooms used in the behavioral experiments: high- and low-spaciousness exemplars, both with and without informative objects masked, plus average-spaciousness exemplars with objects unmasked. Note that this experiment sought to directly measure neural

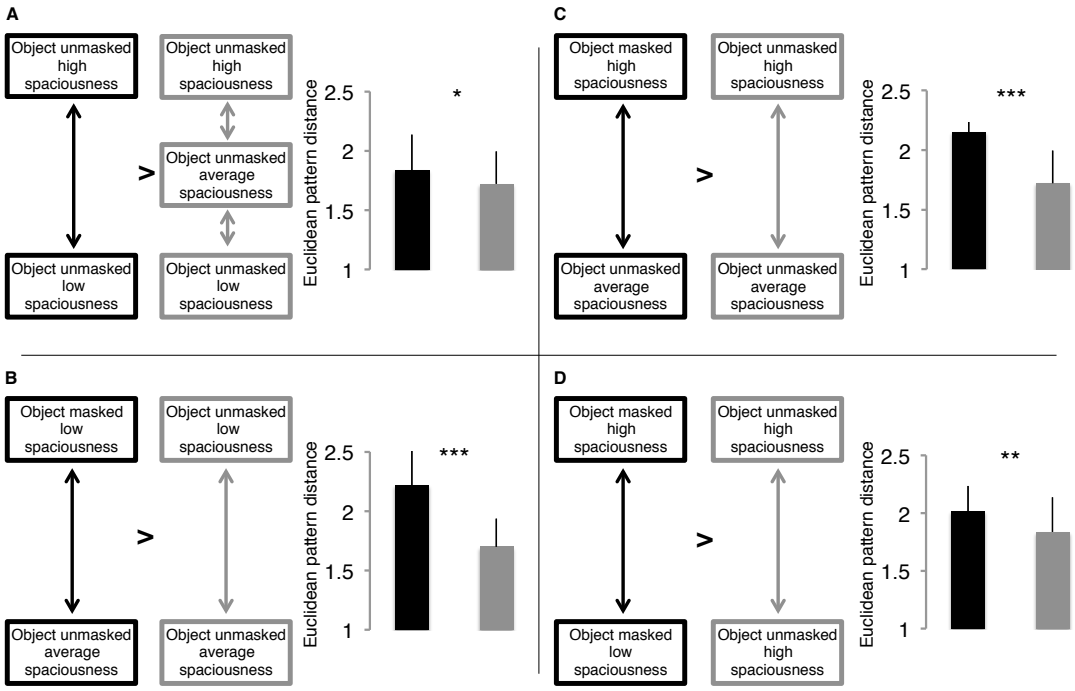


Figure 1.3. Analysis of pattern distances in right PPA. (A) Average Euclidean distances between patterns evoked by high- and low-spaciousness bathrooms with unmasked objects were significantly greater than the average of distances between each of those extremes and the pattern evoked by average-spaciousness bathrooms. (B-D) Pattern distances satisfying the predictions made from behavioral results, demonstrating centripetal object bias in right PPA. The combination of these contrasts was not found in other ROIs. Distance data in each panel correspond to comparisons between patterns denoted by same-shaded arrows in the left half of each panel. Error bars are s.e.m. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001.< 0.01$; ***, $p < 0.001$.

responses to scenes varying in spaciousness and masking state rather than any neural signature of aftereffects those scenes produced. This direct approach was feasible because no perceptual judgments of scene spatial properties were required of subjects in the scanner, who solely judged whether scenes were shaded in gray or blue. Only bathrooms

were used because they were associated with a more reliable centripetal bias in the perceptual experiments.

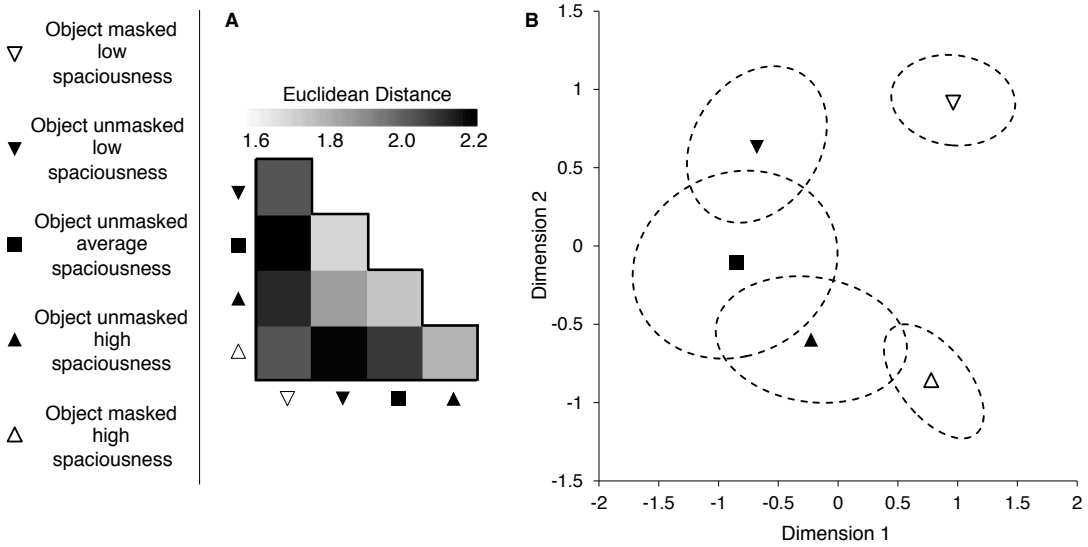


Figure 1.4. Visualization of relationships among scene-evoked patterns in right PPA. (A) Matrix of Euclidean distances among bathroom-evoked patterns, averaged across participants. (B) Corresponding positions of patterns in two-dimensional space returned by MDS; dimensions 1 and 2 capture 36.0% and 30.7%, respectively, of total between-pattern distance. Dashed contours are bootstrap 95% confidence ellipses for pattern coordinates, based on 10,000 resamples. Positions of patterns along the second (vertical) dimension qualitatively match relative encoded spaciousness of scenes indicated by behavioral experiments.

Our analysis focused on the PPA, within which activity patterns have been shown to track the spatial properties of scenes [104,107,129]. Consistent with these previous studies, distances among activity patterns evoked in right PPA by high-, low-, and average-spaciousness bathrooms, all with objects unmasked, qualitatively matched differences among their spatial properties: the average distance between patterns evoked

by high- and low-spaciousness rooms was significantly greater than the average of distances between each of those patterns and patterns evoked by average-spaciousness rooms (Figure 1.3.A, permutation test $p = 0.028$). This "basic" sensitivity to spatial properties was not significant in left PPA, consistent with previous results suggesting greater sensitivity to spatial properties in right PPA [130–132].

Patterns evoked in right PPA by both high- and low-spaciousness bathrooms were significantly closer to patterns evoked by average-spaciousness bathrooms when informative objects in the two extreme scenes were unmasked versus when they were masked (Figure 1.3.B, 1.3.C; $p < 0.0001$ for each). This combination of pattern distances matches the combination of distances among encoded spatial properties that was revealed by our adaptation results, in which scenes with objects unmasked were encoded as more similar to their category average than scenes with objects masked. However, the greater similarity of unmasked extreme scenes to average-spaciousness scenes in fMRI patterns could have also reflected a simple effect of masking state *per se*, rather than an effect of masking on encoded spatial properties. To assess whether this was the case, we also compared similarities between patterns evoked by high- and low-spaciousness exemplars that had objects unmasked to similarities between patterns evoked by the same scenes when they had objects masked. If the greater similarity between patterns evoked by average-spaciousness scenes to those evoked by unmasked high- and low-spaciousness scenes was simply an outcome of object masking by itself, patterns evoked by high- and low-spaciousness scenes should be equally similar to *each other* when objects were masked and unmasked, i.e. when masking state was controlled. Instead, we find that patterns evoked by high- and low-spaciousness scenes were significantly more similar to

each other when objects were unmasked than when masked (Figure 1.3.D, $p = 0.0022$). Thus patterns evoked spatially extreme exemplars with objects unmasked were not only more similar to patterns evoked by average exemplars, but also more similar to those evoked by scenes at the opposite spatial pole. This combination exactly matches the relationships among encoded spatial properties inferred from our perceptual experiments.

Although these results were encouraging, it was possible that the greater distances between patterns evoked by extreme scenes with objects masked arose from differences in cognitive processes related to object masking. Therefore there was still a risk that the greater similarity of patterns evoked by unmasked extreme scenes to patterns evoked by average spaciousness scenes reflected a direct effect of object masking state, rather than an effect of objects on encoded spatial properties. To achieve a more direct comparison between our behavioral results and PPA activity patterns, we used multidimensional scaling (MDS) to visualize and isolate PPA pattern dimensions that specifically corresponded to scenes' spatial properties. Matrices of pairwise Euclidean distances among the five scene-evoked patterns from right PPA (Figure 1.4.A) were averaged across participants and passed to MDS, which produced as output the coordinates of patterns along the set of orthogonal dimensions that best accounted for the full suite of pairwise distances.

The positions of right PPA patterns expressed in terms of the first two dimensions returned by MDS are shown in Figure 1.4.B. Together, these two dimensions accounted for the majority of total pairwise pattern distance, and individually accounted for similar shares of distance (36.0% for Dimension 1 and 30.7% for Dimension 2); the remaining

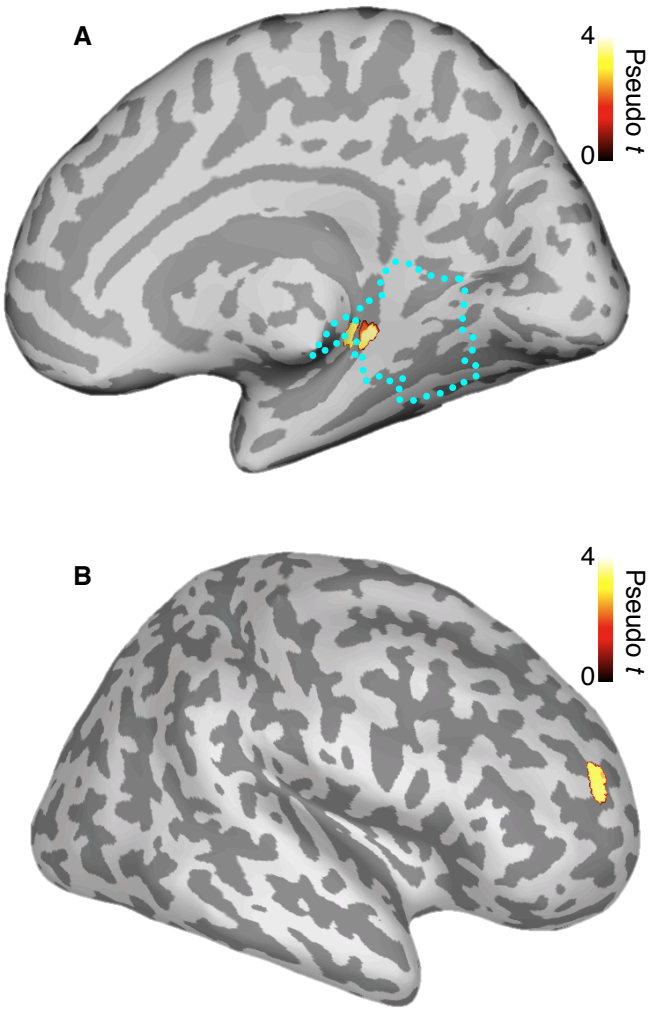


Figure 1.5. Results of searchlight analysis, showing results of second-level analysis of maps of the contrast [(*pattern distance between high/low spaciousness masked bathrooms*) minus (*pattern distance between high/low spaciousness unmasked bathrooms*)]. Statistical thresholds were determined via permutation testing, with voxel activations thresholded at $p < 0.001$ and a minimum cluster size of 7 voxels, which defined the 95th percentile of maximal cluster sizes across 10,000 condition label permutations. (A) A cluster of 8 voxels, centered at MNI 27/-34/-5, fell within the bounds of our group-localizer for right PPA (outlined in blue). (B) An additional cluster of 9 voxels was found in right frontal lobe, centered at MNI 21/47/13.

two dimensions each accounted for substantially less distance (17.5% for Dimension 3 and 15.8% for Dimension 4). Dimension 1 (horizontal axis in Figure 1.4.B) appears to arrange patterns on the basis of masking state, legitimizing our concern that greater distances from patterns evoked by average-spaciousness scenes to those evoked by extreme scenes with objects masked versus unmasked might reflect object contents *per se* rather than spatial property coding. In contrast, dimension 2 (shown vertically in Figure 1.4.B) clearly arranges patterns in order of their evoking scenes' "ground-truth" spaciousness. The coordinate for average-spaciousness scenes along this dimension is roughly intermediate between the coordinates for high- and low-spaciousness unmasked scenes *and* roughly intermediate between the coordinates for high- and low-spaciousness masked scenes. Furthermore, coordinates for both masked and unmasked scenes at each extreme fall on the same side of the coordinate for the average spaciousness scene. These features identify this dimension as capturing PPA sensitivity to scenes' spatial properties [104,107,129]. Neither of the two higher dimensions shared these features (Appendix A Figure 4).

Critically, along dimension 2, coordinates of masked high- and low-spaciousness scenes are further from those of the average scene than are coordinates for their unmasked counterparts, exactly consistent with the centripetal bias we observed in our behavioral results. To assess the probability of observing this effect by chance, we performed MDS on new distance matrices that were computed after random within-subject label swaps between activity patterns elicited by *masked* high-spaciousness scenes and by *unmasked* high-spaciousness scenes, and simultaneously between activation patterns elicited by *masked* low-spaciousness scenes and by *unmasked* low-

spaciousness scenes. Across 10,000 sets of swaps, there was a probability of 0.019 of observing an MDS output dimension which 1) correctly ordered all 5 scenes in terms of their "ground-truth" spaciousness (as described in the previous paragraph) and 2) showed a mask dependent increase in average distance from average- to extreme-spaciousness exemplars that was at least as large as the increase along the second dimension of Figure 1.4.B. (At least one dimension that correctly ordered coordinates was observed for every swap; in the event that two such dimensions were returned, only the dimension accounting for the greater portion of pattern distance was considered.) These analyses indicate that patterns evoked in PPA by extreme bathrooms with objects unmasked were more similar to patterns evoked by average bathrooms specifically along PPA pattern dimensions encoding scenes' spatial properties. No other ROI possessed a profile of pattern similarity consistent with the perceptual experiments. Data from all other ROIs, including the retrosplenial complex (RSC), transverse occipital sulcus (TOS), and lateral occipital complex (LOC), can be found in Appendix A Figure 5 – 15.

Finally, we used a searchlight analysis to identify any regions outside our selected ROIs in which relationships among scene-evoked patterns were consistent with a centripetal bias by informative objects. Consistent with our ROI analysis, in occipitotemporal cortex only voxel clusters corresponding to the anterior portions of PPA showed evidence of centripetal bias (Figure 1.5.A). Evidence for centripetal bias was also found in a single right hemisphere frontal lobe cluster (Figure 1.5.B).

1.4 DISCUSSION

We find that scenes' encoded spatial properties are influenced by the presence of informative objects, which bias encoded properties towards the average of each scene's category. This centripetal bias was evident both perceptually and in activity patterns in PPA, a region that has been linked to processing of scenes spatial properties. Because scenes' *actual* objective spatial properties are to some extent determined by the objects within them, it would not have been surprising to find that the addition of objects exerted a negative effect on scenes' encoded spatial properties. Critically, however, we found that the presence of informative objects led *both* to high-spaciousness scenes being encoded as smaller *and* low-spaciousness scenes being encoded as larger. This contingent directionality indicates that the presence of objects influenced scenes' encoded spatial properties above and beyond what would be expected from objects' simple occupancy of space.

1.4.1 Potential Explanations for Centripetal Bias

Perhaps the simplest explanation for the centripetal bias is feedback from conscious scene category decisions. Assuming that such decisions can drive scenes' encoded spatial properties towards the average of the adjudged category, and further that the presence of objects in scenes leads to more accurate recognition, it stands to reason that a greater proportion of adapting scenes would have been encoded as spatially "average" when informative objects were present. While this can theoretically explain the centripetal bias, it fails practically on several counts. First, category decisions were never required in

either the perceptual or fMRI experiments, and a majority of participants (including all in the fMRI study) viewed scenes from a single category, leaving no impetus for even latent categorization. Second, even though participants in the masked cross-category group did see scenes from both categories, and therefore might have had greater opportunity to categorize scenes, their centripetal bias was no stronger. Third, and most important, the design of our fMRI experiment, in which scenes with and without objects masked were interleaved, meant that the category of masked scenes was always obvious; this applied to our perceptual experiments as well, albeit on a coarser time scale. Thus, even if participants persisted in categorization in the absence of any external motivation to do so, or if such categorization were automatic, it is highly unlikely that categorization accuracy would have differed appreciably between scenes with and without objects masked. Note that this does not challenge our designation of masked objects as "informative", as this designation was based on the frequency of their association with a scene category rather than their impact on categorization in this experiment, although such an impact has been demonstrated previously for these exact object categories [81]. It simply means that, in this particular experimental context, those objects provided no more information about scene category than was already available from other cues.

Two other potential explanations for our results lie in differential attentional demands potentially placed by masked and unmasked scenes. Under both explanations, object masking led to greater attention to scenes' spatial properties and a consequent improvement in the accuracy of their encoding, although in completely opposite ways. In the first, object masking potentially *freed* attentional resources ordinarily attracted by objects to be deployed to adapting scenes' spatial properties, which were therefore

encoded with greater fidelity to scenes' true extreme spatial scales than when objects were unmasked. To accept this explanation, however, one must adopt the general view that codes for spatial properties are inherently less accurate when objects are present. Considering that almost all real-world scenes contain objects, this inaccuracy would be maladaptive, and therefore unlikely. In the second attention-based explanation, object masking drew greater attention to masked adapting scenes *as wholes* as observers struggled to identify masked objects, with the outcome again that the encoded values of these scenes' spatial properties was more faithful to their extreme natures than when objects were unmasked. While this explanation cannot be directly discounted, the rapid timing of adapting sequences is likely to have dampened any efforts by participants to decipher masked objects, particularly given the ongoing repetition detection task. One way to avoid this potential problem could have been to simply excise, rather than mask, informative objects in scenes. Doing so, however, would have introduced other differences between masked and unmasked adapting scenes, including alterations to scenes' low-level statistical properties. We wished to avoid such differences for the sake of our subsequent fMRI experiment, given evidence that PPA is sensitive to them [133].

Moving beyond attention, it is very difficult to explain the centripetal bias as an outcome of some *direct* cognitive effect of object masking (i.e., an effect not mediated by some change in encoded spatial properties). Although it is possible and perhaps even likely that cognitive processes related to object recognition were differentially activated by masked and unmasked adapting scenes, this difference is unlikely to explain our results, for two reasons. First, it is unlikely that purely *object*-related cognitive differences would have influenced aftereffects exerted on the perceived *spatial* properties

of test scenes. Second, even if they were able to exert such an influence, it is even less likely that they could have produced the *bidirectional* enhancements in aftereffects we observed. This is because any object-based differences in cognitive processes between masked and unmasked adapting scenes would have been identical for both high- and low-spaciousness adapters, and as a consequence any potential contamination of spatial codes should therefore have likewise been identical. This conflicts with our observation that object masking exerted opposite effects on the encoded spatial properties of high- and low-spaciousness adapting scenes.

Finally, it is unlikely that the differing strength of aftereffects we observed with masked and unmasked adapters reflects the fact that test scenes "matched" the masking state of unmasked adapters but not of masked adapters. We acknowledge that, in general, the susceptibility of a test stimulus to aftereffects is dependent upon the degree to which it matches the adapting stimulus along non-adapted dimensions. For instance, face-specific aftereffects are stronger when the adapting stimuli and judged stimuli occupy the same retinal location [134] and motion aftereffects are strongest when adapting and judged stimuli possess the same spatial frequency [135]. However, if we liken the masking states of adapting scenes in our study to different retinal positions or spatial frequencies, aftereffects should have been *stronger* for unmasked scenes because they matched the masking state of test scenes. This, of course, is the opposite of what we observed.

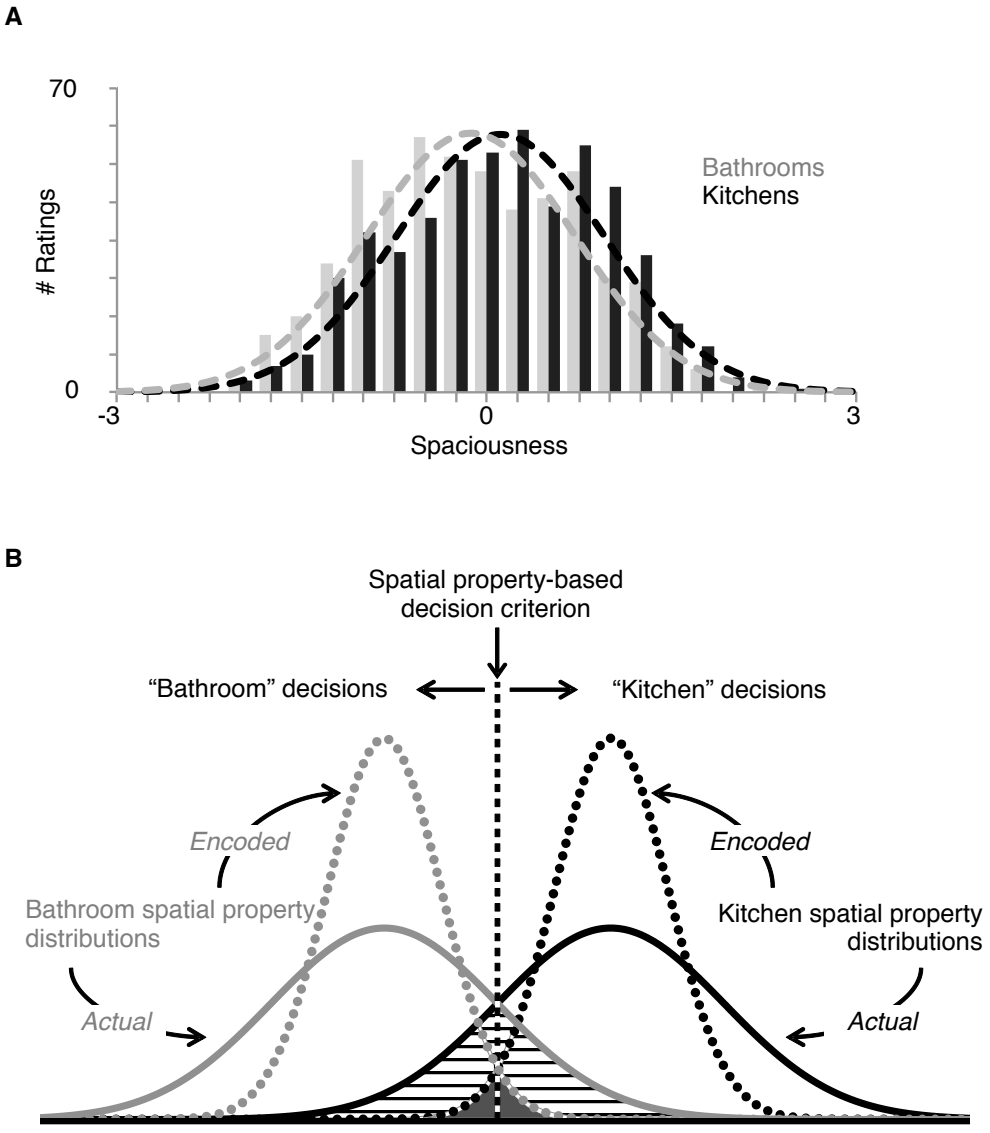
"Crosstalk" theory

Rather than an outcome of decision feedback or attention, we propose instead that the centripetal bias reflects a form of heretofore undescribed "crosstalk" between object-

and spatial property-encoding pathways. In this theory, objects associated with a given scene category contribute a "normalizing" signal to

Figure 1.6. Hypothesized role for centripetal bias in scene categorization. (A) Histograms of crowd-sourced ratings of the spaciousness of the 100 bathrooms (gray) and 100 kitchens (black), all with unmasked objects, from the middle spaciousness quintile of each category, accumulated across 61 observers (bathroom $n = 609$, kitchen $n = 607$). Ratings were solicited from paid online raters (separate from those who contributed to scenes' quintile assignments) who were each asked to rank a pool of 50 bathroom and 50 kitchen images in terms of perceived spaciousness of the depicted rooms, without regard to category. X-axis values are within-subject z-transforms of raw ratings. The means of these distributions are significantly different (two-tailed t-test, $t(1214) = 4.93$, $p = 9.3 \times 10^{-7}$). These data are shown only to establish that average-sized real-world bathrooms tend to be judged as smaller than average-sized real-world kitchens, albeit with significant overlap. We infer from these data that average-sized rooms in each category possess similarly differing distributions of *actual* spatial scales. No inferences about scene categorization mechanisms are drawn from these data. Dashed curves are normal distributions fit to data. (B) Schematized version of distributions in A. Because the distributions of spatial properties overlap between the two room categories, any fixed spatial property-based decision criterion (vertical dashed line) will result in some proportion of spatial-property based categorizations which conflict with scenes' object contents; this fraction is represented by the union of the horizontal-lined and dark gray-shaded regions. (This example assumes that object contents are perfectly informative of

scene category.) By narrowing the distributions of encoded spatial properties (dotted curves), object-triggered centripetal bias reduces overlap between the distributions of the internal representations of the categories' spatial properties, potentially producing a smaller proportion of conflicted categorizations (dark gray-shaded region alone). Although centripetal bias is illustrated here as a reduction in the standard deviations of normal distributions, non-uniform centripetal effects (e.g., applied only to distributions' tails) would produce a similar outcome.



codes for spatial properties, bringing potentially large excursions in encoded values into closer register with those typical of the scene category the objects are associated with. In contrast to the feedback account rejected above, in this framework the centripetal influence of objects *precedes* scene recognition. Moreover, we propose that the purpose of this influence is to *assist* scene recognition by easing potential conflicts between scene category judgments derived from object contents and spatial properties.

For an example of how this might work, let us return to the task, described in the introduction, of deciding whether a room in an unfamiliar house is a bathroom, perhaps after being told that both a bathroom and a kitchen (and no other room type) can be found along a hallway one is walking down. These room categories differ both in their object contents (Table 1) and in their average spatial properties (Figure 6A). As such, upon viewing the first encountered room it can be expected that hypotheses about its identity could be generated from both its object contents and its spatial properties. (We use the term "hypothesis" here to avoid any mechanistic implications attached to "schema" or "context frame".) Let us assume that this room happens to be an inordinately large bathroom. Owing to the high degree of overlap between real-world distributions of the sizes of bathrooms and kitchens, it is quite possible that this room's extreme spatial properties may place it on the "kitchen" side of a neutral spatial property-based category criterion, generating a spatial property-based hypothesis that conflicts with the hypothesis generated from its object contents, which we will assume for present purposes are fully dispositive of category. A final judgment of the room's identity therefore requires some means of reconciling these competing hypotheses. This in turn requires consideration of

variety of factors to determine the appropriate weight that should be given to each hypothesis, a process that might take time and offer added opportunities for error.

We propose that the object-triggered centripetal bias we observed aids scene recognition by reducing the frequency with which this reconciliation process is required. By driving the encoded spatial properties of the very large bathroom towards those of the average bathroom, centripetal bias reduces the probability that the hypothesis of scene identity derived from those properties will conflict with the hypothesis derived from the scene's object contents. Assessed across encounters with many scenes, we propose that centripetal bias narrows the distributions of each scene category's *encoded* spatial properties, reducing the degree of overlap between categories relative to their distributions of *actual* spatial properties and consequently decreasing the proportion of scenes on the "wrong" side of a neutral spatial property criteria between categories (Figure 1.6.B). We expect that the resulting harmonization of category hypotheses derived from encoded spatial properties with those derived from objects would improve the speed and accuracy of categorization.

Based on this theory, we expect that the degree of centripetal bias produced by an object should bear some relationship to the strength of its association with a specific scene category; that is, that the identities of the masked objects in both our behavioral and fMRI experiments mattered. Although our study did not directly test this relationship, the alternative that all objects are equipotent in inducing centripetal bias is virtually impossible to reconcile with the bidirectional bias we observed. Consider a completely empty room with a floor area intermediate between the average floor areas of two room categories generally differing reliably in size. The addition of an object with no

association to any particular scene category will, by definition, add no information about the identity of the scene, leaving the direction of any potential induced bias unspecified: should the bias be towards the smaller or the larger scene category? With the target of bias undefined, such an object cannot produce *any* bias, negating the original assertion that objects are equipotent in producing the bias. We therefore consider it extremely likely that the centripetal bias we observed depended on the identities of those objects which varied in masking state.

We acknowledge, however, that our results do not tell us whether the ability of objects to bias scenes' encoded spatial properties derives from objects' statistical associations with specific base-level scene categories, such as "bathroom" versus "kitchen", or with scenes grouped at some higher taxonomic level, such as "indoor scenes" versus "outdoor scenes". In other words, while our results are consistent with our hypothesis that objects bias encoded spatial properties towards the average values of bathrooms or kitchens, they leave open the possibility that objects biased encoded properties towards those of the average indoor room. This ambiguity exists because the high- and low-spaciousness adapting scenes we used from each scene category were possibly extreme enough that they bracketed the average spatial properties of all indoor rooms, not just the average of their own category. That is, high- and low-spaciousness bathrooms exemplars in our study were likely so extreme that they were likely larger and smaller, respectively, than not only the average bathroom, but also the average kitchen, as well as the averages of several other common room categories. However, while we cannot identify with certainty the level of scene specificity at which the centripetal bias operated, it seems that a bias which targeted the spatial properties of specific categories

would be more adaptive than one which targeted the average properties of a higher taxonomic cluster, such as indoor scenes. This is because while a bias targeting the average indoor room would benefit indoor/outdoor scene distinctions, it would simultaneously harm distinctions among base-level categories of indoor or outdoor scenes by compressing the range of encoded spatial properties of all categories within each group towards a single point. In contrast, a bias that targeted base-level scene categories, and therefore aided distinctions among them, would be at worst neutral with respect to the high taxonomic distinctions such as indoor versus outdoor. Ultimately, additional experiments are necessary to clarify this issue. We emphasize, however, that the uncertainty we highlight does not challenge our crosstalk interpretation of the centripetal bias; it merely raises questions about the taxonomic level of scene distinctions that might benefit from crosstalk: base-level or superordinate.

Although we favor the idea that centripetal bias targets base-level scene categories, our crosstalk theory does not predict that all scenes will be equally susceptible. Instead, assuming equally strong object associations, centripetal bias should vary in strength with the strength of scene categories' associations with any particular set of spatial properties. Specifically, we predict that centripetal bias should be stronger for indoor scenes (such as the bathrooms and kitchens used here), which tend to occupy a relatively narrow range of real-world sizes, than for outdoor scenes. Given this, we do not interpret the fact that objects "controlled" spatial properties in this experiment to indicate that objects enjoy a general position of superiority over spatial properties during processing of all scenes. Thus, an important future test of our crosstalk hypothesis will be to show not only that centripetal bias exists beyond the narrow range of scene types used

in the present study, but also that it fails predictably for scene categories not strongly associated with any particular spatial scale.

We wish to emphasize that our crosstalk theory is not simply a restatement of the idea that objects activate scene schemata or context frames storing information about the features, including spatial properties, typically associated with each scene category [99]. Instead, we conceive of crosstalk as a direct *translation* of object information into spatial property codes, independent of (although perhaps coincident with) schemata activation. Support for this view comes from the fact, already mentioned, that the identities of masked adapting scenes were likely to have been quite obvious to participants, whether due to remaining identifying features, or, for nearly half of our participant pool, previous exposure to adapting scenes in their unmasked forms as well as intervening test scenes. This makes it likely that scene schemata were equivalently activated regardless of adapting scenes' masking states. The persistence of centripetal bias in light of this suggests that object processing pathways enjoy direct access to spatial property codes that is distinct from their capacity to activate scene schemata. Similarly, it suggests that the influence of objects on encoded spatial properties arises from the objects' *visual features*, rather than their *identities*, since we expect that the latter also remained fairly firmly instantiated from context even during masked blocks.

Our crosstalk theory thus holds that informative objects benefit scene categorization in two ways: by directly activating schemata of their associated scenes, and by biasing encoded spatial properties to reduce conflicts with properties associated with those schemata. This view makes the testable prediction that the presence of informative objects should aid performance on a binary scene discrimination task *more*

under conditions that allow objects to produce a centripetal bias, such as when scene exemplars possess spatial properties departing from their category averages, versus when they do not, such as when exemplars from at least one category already match the spatial properties typical of their category. The competing view that the centripetal bias reflects feedback from object-activated schemata predicts no such difference. We expect, therefore, that future experiments will be able to clarify whether our feedforward "crosstalk" explanation of centripetal bias is correct.

1.4.2 Role of Parahippocampal Cortex

Matching our behavioral results, activity patterns evoked in right PPA by scenes at each spatial extreme were quantitatively more similar to patterns associated with the opposite extreme when objects were unmasked versus when they were masked, both in general and specifically along a pattern dimension which correctly ordered scenes' actual spatial properties. This correspondence with our behavioral results was not observed in any other ROI. Although PPA has been shown to be sensitive to low-level properties of stimuli, such as spatial frequency [133,136] and texture [137,138], response differences between unmasked and masked scenes are unlikely to have arisen from differences in low-level properties: any influence of object masking on these properties should have taken the same sign for both high- and low-spaciousness scenes, whereas the influence of objects along the PPA space-coding dimension operated in opposite directions depending upon scene spaciousness. While PPA activity has been shown previously to relate to spatial properties of scenes [104,107,129] and to human judgments of scene category [139,140], the present study joins a very small group demonstrating that PPA activity tracks scenes'

encoded spatial properties even when those properties depart from physical reality [141,142].

Viewed in the framework of our crosstalk theory, our results suggest that PPA, at least in the right hemisphere, is the brain area in which encoded spatial properties of scenes are brought into alignment with expectations derived from scenes' object contents. Our assignment to PPA of this role as junction point between codes for objects and spatial properties is consistent with its recent characterization [129] as the midpoint in a hierarchy of scene processing regions which ranges from the purely object-sensitive LOC to the purely space-sensitive retrosplenial complex (RSC). Moreover, our results offer an alternative perspective on the contentious issue of the origin of object-evoked activity in PPA [108,129,143,144], which has alternately been explained in terms of either object-triggered spatial representations [105,106,145] or contextual associations among objects [108,146,147]. Our results suggest that the presence of object-evoked activity in PPA also reflects the object information necessary for centripetal bias to take place. Indeed, as our crosstalk theory is based on associations between spatial properties and non-spatial information (i.e., object identity), the role we ascribe to PPA as the effector of centripetal bias appears consistent with both the context- and layout-centered views of its function in scene processing.

Neither our ROI-level nor searchlight analyses showed a similar effect of objects on encoded spatial properties in left PPA. This laterality can be separated into two distinctions between right and left PPA. First, unlike patterns from right PPA, patterns from left PPA failed to pass even the basic test of distinguishing significantly among unmasked scenes on the basis of spatial properties: pattern distances between high- and

low-spaciousness unmasked exemplars were not significantly greater than pattern distances between those exemplars and the average-spaciousness exemplars. The reason for this failure is not clear, although some research has suggested that left parahippocampal cortex may have a relatively reduced capacity for spatial processing [130–132,148], which may have been less apparent in previous MVPA studies of PPA spatial sensitivity that used scenes spanning a much greater range of spatial properties than those we used [104]. Second, we observed no significant effect of object masking on relationships among left PPA activity patterns. This potentially reflects the demonstrated greater sensitivity of right parahippocampal cortex to the specific visual contents of scenes, contrasting with a greater capacity for abstraction in left parahippocampal cortex [148–150].

While the medial temporal cluster identified by our searchlight analysis fell within the boundaries of group-defined PPA, it is positioned markedly anteriorly in parahippocampal cortex. Our results thus join a growing set of findings suggesting that PPA is differentiable along its rostrocaudal axis in terms of both response properties [133,151] and connectivity [152], and dovetails very closely with some. For instance, anterior PPA has been recently shown to be much less sensitive to objects than posterior PPA [152], and less sensitive to the high spatial frequencies that might convey information about objects [133]. While this might appear at first to conflict with our searchlight map showing the most prominent effect of objects in anterior PPA, it is important to remember that the searchlight analysis identified subregions whose patterns were *biased* by the presence of objects in scenes, not necessarily those which carried information about the identities of the objects. Furthermore, inasmuch as the centripetal

bias would appear to be a rather high-level refinement of spatial property codes, it makes sense that it would be found in anterior PPA, which shares a greater degree of connectivity with fronto-parietal networks than posterior PPA [152]. It is noteworthy in this regard that the only other area which showed evidence of centripetal bias in our searchlight analysis was a cluster in prefrontal cortex (PFC). Whether this indicates some functional association with PPA is unclear, but to the extent that it might, we are inclined towards the view that it results from prefrontal mirroring of a centripetal bias that arises in PPA, potentially reflecting the channel through which PPA spatial codes contribute to categorical decisions.

In summary, although scenes' spatial properties and object contents are formally independent descriptors of scenes, both our behavioral and fMRI results show that this theoretical independence is not respected by the visual system. While it has long been appreciated that objects can influence judgments of scene category, the biasing influence of objects on encoded spatial properties that we observed has not been previously described, nor explicitly predicted by scene recognition models. We propose that this bias reflects a system of object/spatial property crosstalk supporting generation of unified judgments of scene category by reducing potential categorization conflicts. Further perceptual and neuroimaging experiments will be necessary to understand the neuroanatomical basis of this phenomenon, and to explicitly test the hypothesis that it aids the accuracy and speed of scene recognition.

2.0 OBJECT PERCEPTION DURING SCENE CATEGORIZATION IS INFLUENCED BY SPATIAL PROPERTY ASSOCIATIONS

Manuscript ***in preparation***

The ability to categorize a scene depends on complementary resources of visual information, describing both the objects and spatial properties (e.g. spatial layout) of scenes. Although standard theory considers these resources to be processed essentially independently during scene categorization, Chapter 1 provides evidence that they are combined at a perceptual level, evident in a systematic bias of spatial property perception by objects strongly associated with scenes. However, little else is known about this perceptual combination of scene information. Among open questions are: how is this combination instantiated in the visual system, whether it only biases perceived spatial properties or operates bidirectionally between objects and spatial properties, and whether it actually impacts scene categorization accuracy, as we have theorized.

In this chapter we demonstrate that implicitly learned statistics linking co-occurring object and spatial property information are drawn upon during scene categorization to perceptually "fill-in" obscured object information in scenes. The ability of scenes' spatial properties to bias their perceived objects is robust, replicating over multiple experiments that control for a variety of potential confounds and evident in significantly more accurate scene categorization. This set of experiments validates the theory of a perceptual combination of scenes' object and spatial property information during scene categorization and identifies its mechanistic underpinnings.

2.1 INTRODUCTION

For healthy humans, everyday behavior depends on navigating effectively through the world. But in order to navigate, humans must first effectively recognize their

surroundings. While this at times requires identifying a specific scene, such as when recognizing our own homes, we are more often faced with the task of determining a scene's category. For instance, when visiting a museum one must distinguish between the gift shop and the gallery. Fast and accurate scene classification is a hallmark of the human visual system, which depends on complementary visual features [14,116]: 1) information about scenes' objects, such as their identities and spatial configuration [1–6,153], and 2) scenes' intrinsic spatial properties, such as size and expanse [8–10,15,18,19,154–157]. While humans can successfully categorize scenes based on either of these resources when experimentally isolated [5,116], real-world scenes are typically categorized based on their combination [5,98].

Theory has long held that scene categorization unfolds in two stages. In the first, visual features describing scenes' objects and spatial properties are independently processed in the visual system. This independence is maintained until the second stage of processing, taking place downstream in regions of cortex involved in decision-making, where scene category is read-out from some weighted combination of these resources [75]. However, evidence presented in Chapter 1 for a systematic, "centripetal bias" in the perception of indoor scene spatial properties challenges this theory, suggesting instead that these resources are entwined during the first stage [35]. We found that the perceived spatial properties of indoor scenes are biased towards the spatial properties associated with the identities of those scenes' objects. In other words, the presence of an oven within a scene biased its perceived spatial properties towards that of the average kitchen. This observation revealed that the visual system has access to information about the spatial

scales typically associated with objects, which influences how spatial properties are perceived.

But how exactly are these associations instantiated? The most straightforward explanation is that the visual system tracks the statistical co-occurrence of particular objects with particular scales of spatial properties – analogous to a perceptual look-up table in which values in "object space" are associated with values in "spatial property space". With such a look-up table in place, activation of neural codes for certain objects could activate neural codes for specific sets of spatial properties, resulting in centripetally biased spatial property perception. Lending plausibility to this model are studies demonstrating that humans are implicitly sensitive to regularities in visual information [158–160], learning statistics that capture both spatial and temporal features of objects, which improve subsequent perception. For instance, synthetic objects are more accurately recognized when they are displayed either in an 2-Dimensional array or a temporal sequence of other familiar synthetic objects [161–165].

Nevertheless, little is known about the impact of object and spatial property co-occurrence statistics on scene categorization. We reasoned that if the visual system does have access to these statistics they should not only impact perception of scenes' spatial properties, as demonstrated in Chapter 1, but also permit spatial properties to bias the perceived identities of objects in scenes. Evidence for biases in both object and spatial property directions would indicate that the visual system leverages co-occurrence statistics during scene categorization to reinforce and align these resources before decision-making. Through this influence these co-occurrence statistics could improve scene categorization accuracy by facilitating downstream read-out of perceptual

information during decision-making, particularly when viewing a scene with obscured objects or extreme spatial properties.

Here we explored these questions by asking if implicitly learned statistics describing scenes' co-occurring objects and spatial property features bias perception of their objects during categorization. To do this, we compared the performance of two groups of participants at categorizing scene exemplars belonging to novel scene categories. Importantly, only one group was given the opportunity to learn co-occurrence statistics during an initial training phase. During the subsequent categorization phase, both groups saw scenes that had perceptual masks obscuring their object contents, leaving spatial properties as the only resource available to both groups' category decisions [15]. We hypothesized that the group given the opportunity to learn co-occurrence statistics would, when viewing scenes with objects masked, preconsciously "fill in" the masked objects based on the scenes' spatial properties. Having information in both their object- and spatial property-processing pathways, these participants would be significantly more accurate during scene categorization than participants who had never learned co-occurrence statistics. This approach to assessing the influence of spatial properties on object identity was preferable to a task in which participants were asked to name objects because it avoided the possibility that improvements in object naming might be due to feedback from mature scene category decisions.

We also explored how object and spatial property co-occurrence statistics are stored. We reasoned that the amount of information needed to form a complete mapping between all possible combinations of scenes' objects and spatial properties is too large to tractably store in the human visual system [166]. Instead, we propose that an efficient

solution is to maintain "parameters" that establish an intermediate mapping: capturing co-occurrences between *commonly encountered* object and spatial property features. In this parameterized account, the perceived features from information resource (e.g. spatial properties) are matched to their most similar intermediate parameters held in memory (e.g. "small room"), which supports an easy read out of the commonly associated values from the other perceived resource (e.g. sink). For instance, this means that *any* kitchen-sized scene will be influenced by a spatial property parameter encouraging the observer to perceive it as containing an oven. This account predicts that co-occurrence statistics learned for some scenes can generalize their influence and improve categorization accuracy of novel scenes, for which co-occurrence statistics had not been learned.

We found evidence that scenes' spatial properties influence the perception of their object features (Experiment 1). This influence operates off of implicitly learned statistics describing scenes' co-occurring object and spatial property features, and leads to significant gains in scene categorization accuracy. In two replications, we found that this influence does not depend on participant attention (Experiment 2) or temporal regularities (Experiment 3) during co-occurrence statistic learning. We also found that learning these statistics for some scenes facilitated recognition of other, similarly sized scenes for which they were not learned, indicating an efficient and parameterized storage (Experiment 4).

2.2 EXPERIMENT 1

2.2.1 Materials and Methods

2.2.1.1 Participants

A total of 112 participants (63 males) between 19-65 years old were recruited for the experiment from Amazon Mechanical Turk (AMT; https://www.mturk.com/), an online service for web-based experiments. All participants were based in the United States, provided written informed consent in accordance with the procedures of the Boston College Institutional Review Board, and were paid $1.50 for participation, lasting approximately 15 minutes.

2.2.1.2 Stimuli

Visual stimuli were computer-generated exemplars of two "novel" scene categories. Novel categories were generated by combining objects and scene spatial properties that are not reliably observed together in the real world [167]. Novel rather than familiar scene categories were used to directly manipulate how participants learned object and spatial property co-occurrence statistics.

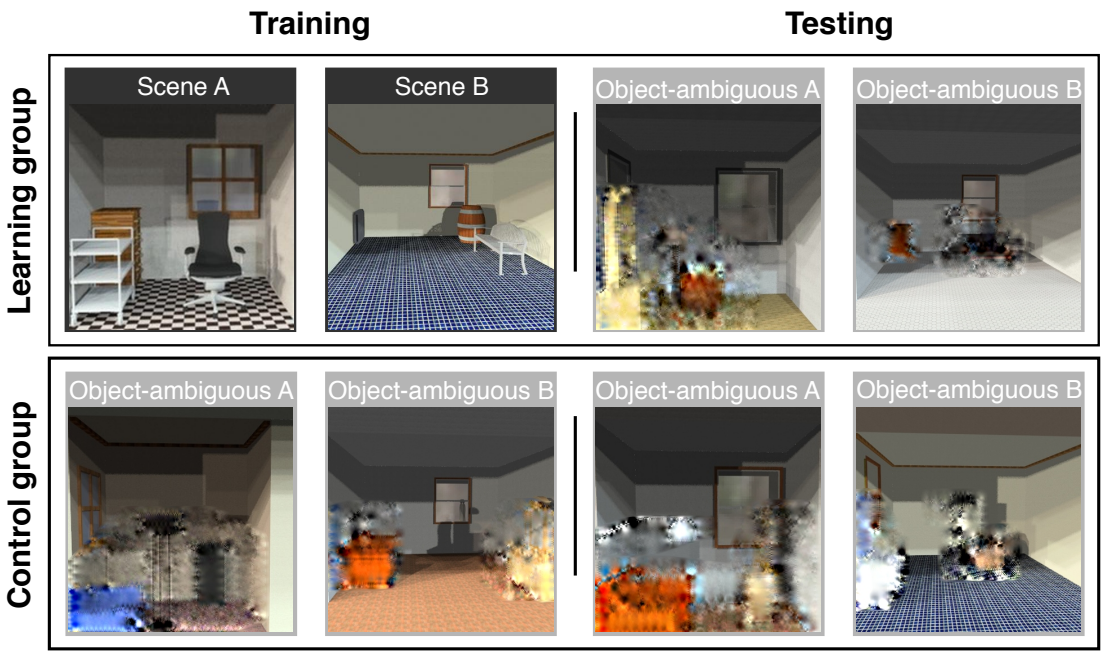


Figure 2.1. Experiment stimuli. Scene A contained a chair, cart, and dresser and was smaller than Scene B, which contained a bench, radiator, and barrel. During training, the Learning group viewed object-intact versions of these scenes, whereas the control group viewed versions that were object-ambiguous. During testing, scenes were object-ambiguous for both groups.

The two novel scene categories ("Scene A" and "Scene B", Figure 1) were constructed by assigning distinct ensembles of objects to indoor rooms drawn from two adjacent and non-overlapping ranges of floor area in order to establish different spatial properties for the two categories. Object ensembles occupied similar spatial extents within exemplars for each category. Each Scene A exemplar contained a chair, cart, and dresser, whereas each Scene B exemplar contained a bench, radiator, and barrel. Object models in each scene image were randomly drawn from a pool of 10 for each type of object, and positioned in one of 24 randomly selected configurations.

Each scene category contained 150 unique exemplars. Floor areas ranged in 0.035 square meter increments from 4.5 square meters to 9.73 square meters for Scene A and 9.77 square meters to 15 square meters for Scene B. Rooms were rendered using Trimble Sketchup (www.sketchup.com), IRender nXt 4.0 (www.renderplus.com), and custom Ruby scripts. Separate versions of each image were produced in blue and gray scale and cropped to 400 x 400 pixels.

Two versions of each exemplar were used for the experiment: an "object-intact" version, as described above, and an "object-ambiguous" version in which the objects were replaced by ineffable wavelet-scrambled [113] versions of a couch, lamp, and cabinet (i.e., objects not associated with either new category; "Object-ambiguous A" and "Object-ambiguous B", Figure 2.1). The use of perceptually masked objects was preferable to excluding objects from these rooms to maximize their spatial similarity to object-intact exemplars in Scene A and Scene B.

2.2.1.3 Procedure

This experiment used a between-subjects design to investigate how learning the statistical co-occurrence of scene features impacts later categorization of object-ambiguous scenes. One group of participants was given the opportunity to learn object and spatial property co-occurrence statistics of Scene A and Scene B (Figure 2.1, "Learning group"), while the other was not (Figure 2.1, "Control group"). This manipulation was implemented during an initial training phase: the Learning group viewed object-intact exemplars from Scene A and Scene B while the Control group viewed object-ambiguous exemplars. Since the objects in object-ambiguous exemplars carried no signal for scene category, the Control group was unable to learn co-occurrence statistics in these scenes.

The expectation of this experiment was that Learning group participants would learn co-occurrence statistics of the respective object contents and spatial properties of the two scene categories during the training phase through exposure to object-intact scenes. This was accomplished by ordering scene exemplars in the training phase as random walks through the graph depicted in **Figure 2.2.A**. Each node in this graph corresponds to a bin of 30 spatially similar exemplars belonging to one of the scene categories (figure 2.2.A legend), and the configuration of the nodes enforced imbalanced transition probabilities between categories. When the random walk for a stimulus sequence landed on the node denoted by the black square (Figure 2.2.A, "5 "), a participant would see a spatially large exemplar belonging to Scene A that had a 75% chance of being followed by an exemplar from the same category (in comparison to an exemplar drawn from a gray node, which was always followed by one from the same category). These imbalanced transition probabilities during scene viewing provided implicit category boundaries for novel scenes, which we expected would maximize participants' ability to learn co-occurrence statistics. Similar designs have previously been used to promote category boundaries among visually unrelated stimuli [168]. Control group participants saw stimulus sequences arranged according to the same random walk procedure, except these sequences contained scenes with all objects masked so as to be unidentifiable.

The training phase consisted of 154 images displayed for 1500ms each (Figure 2B). In order to maintain attention during the experiments, participants judged each scene exemplar with the + or – key on their keyboard as being blue- or gray-scale while it was on screen (the shading procedure is detailed in the Stimuli section). Participants were

provided feedback for their color judgments during the training phase to encourage engagement: the currently viewed image was outlined green for a correct response or red for an incorrect response.

Following this, participants completed a testing phase that measured the impact of co-occurrence statistic learning on scene categorization (Figure 2.1, Testing). Learning and Control group participants completed the same Testing procedure. Participants were told that there were two scene categories at the start of the testing phase and were asked to categorize object-ambiguous scene exemplars as belonging to Scene A or Scene B (instructions can be found in Appendix B; Figure 2.2.C, Testing). Since the objects in these scenes carried no signal, the task relied on judging scenes' spatial properties.

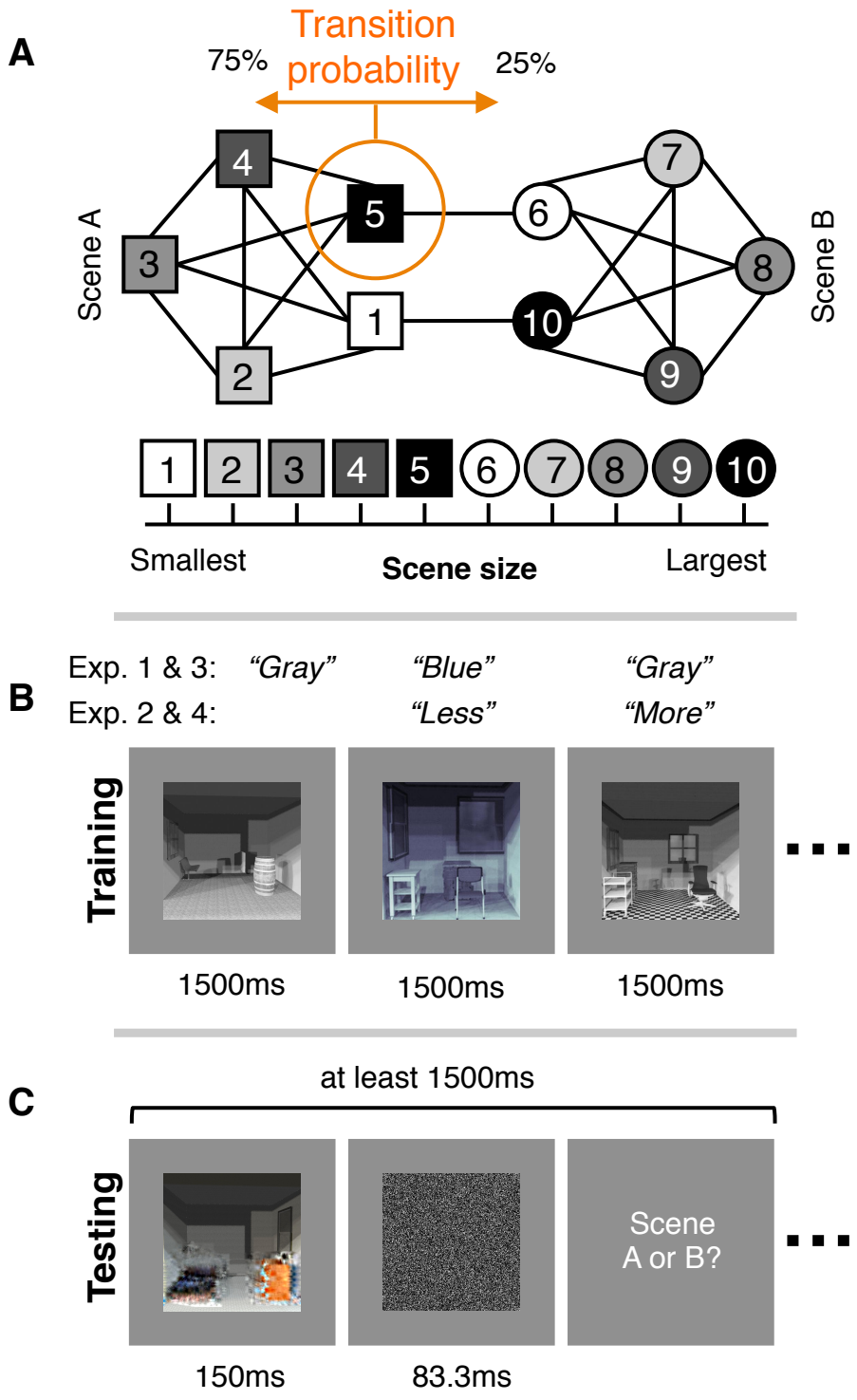
Color images of testing phase scene exemplars were displayed for 150ms, followed by 83.3ms of a white noise mask, and an indefinite prompt to judge if the scene belonged to Scene A or Scene B with the + or the – key on their keyboard. Participants were encouraged to answer as quickly as possible, but each stimulus event lasted at least 1500ms so that each participant had an equal number of judgments. When judgments were entered before 1500ms, a dot was shown at the center of the screen until the remaining time had lapsed. The order of testing phase sequences was uniformly random to reduce the likelihood of temporal regularities learned during the training phase carrying over and influencing decisions. Participants categorized 154 exemplars during the testing phase, however a response to the first image was not recorded due to software error. For each experiment, participant performance was taken as the mean accuracy across these 153 judgments.

2.2.1.4 Experimental Setup

Amazon Mechanical Turk directed participants to the experiments, which were hosted on Psychophysik (http://psk.bc.edu), a web-based application for creating online psychophysics experiments. Psychophysik provided several features that controlled the quality of the data received from participants. These features were high-resolution javascript timers for recording participant reaction times, stimulus display times, and the time of day of participation; extracting information about each participant's operating system and screen resolution; and enforcing full-screen mode during participation. Additionally, the application was restricted to Google Chrome for optimal consistency between participants. Data analyses detailed below calculated the extent to which any of these variables contributed to testing phase classification accuracy. The entire Experiment 1 can be accessed at http://bit.ly/co_psk.

Figure 2.2. Experimental procedure. (A) Training phases in Experiments 1, 2, and 4 were sequenced with a random walk through this graph structure, which biased transition probabilities between scene categories. This promoted statistical learning by making it more likely for the large exemplar from Scene A ("5") to be followed by an exemplar from Scene A than by an exemplar from Scene B. The shape of each node indicates its scene category, and the shade of gray its relative size. (B) During training, exemplars were displayed for 1500ms. Depending on the experiment, participants either judged the color shading of each exemplar (Experiments 1 and 3), or its "spaciousness" relative to the one preceding it (Experiments 2 and 4). (C) During testing, object-ambiguous exemplars were displayed for 150ms, followed by a 83.3ms of a white-noise mask, and

finally a prompt to judge the exemplar's category, which remained on screen for at least 1500ms. Scenes were shown in a completely random sequence during the testing phase.



2.2.1.5 Data Analyses

Our approach for web-based experimentation supported rapid collection of data from a large set of participants. However, we speculated that this approach would be vulnerable to two sources of measurement error: 1) Participant disengagement and distraction, 2) time-of-day effects. To control for these factors in participant data, we employed a set of preplanned and data-driven filters. The filters and their effects on data collection are detailed below. Results are based on data from participants passing the entire ensemble of filters.

2.2.1.6 Participant Disengagement

Given the online nature of these experiments, which does not allow any supervision of participants, we expected that some participants were less engaged than others and entered judgments unrelated to the task at hand in order to simply progress through the experiment. Data from these participants would add noise to our analyses and reduce the ability to detect differences between Learning and Control groups. We therefore adopted an unbiased, data-driven strategy to identify and exclude disengaged participants from analyses.

To measure participant engagement, we took advantage of data collected during each participant's training phase. During training, participants discriminated each exemplar as being blue- or gray-scale, in a task that was designed to increase participant engagement while requiring perceptual processes orthogonal to the main experimental manipulation measured during the testing phase. This task also provided us with data that we used to distinguish between engaged and disengaged participants. Each participant's sequence of training phase responses, evaluated as correct or incorrect, was passed as a

high-dimensional vector to a hierarchical agglomerative clustering algorithm (ward link; however, clusters were relatively robust to the choice of linkage) [35]. This approach grouped together participants with similar response patterns across the training phase, separating those with heterogenous (and high-accuracy) patterns from those with homogenous (and low-accuracy) patterns. We considered the group of 4 (1 Control) participants in the homogenous group to be disengaged from the experiment and excluded them from analysis.

2.2.1.7 Time-of-day Effects

Performance in perceptual experiments has been tied to the time-of-day of participation [169]. We employed a filter to control for these effects, excluding participants who completed the experiment 2 standard deviations above or below the average time-of-day of their experimental group. This filter identified 2 Participants (1 Control) for exclusion from the analyses. In total, of the original 112 participants (54 Control) 105 participants (52 Control) remained in Experiment 1 following both the engagement and time-of-day filters.

2.2.1.8 Statistical Analyses

To support statistical testing, we equalized the number of participants between Learning and Control groups. We did this with a bootstrapping procedure, which estimated the larger group's average performance after randomly resampling a sample equivalent to the size of the smaller group over 10,000 iterations.

Permutation tests measured the statistical significance of differences in classification accuracy between Learning and Control groups in each experiment. For

each experiment, we randomized the group membership label (i.e. Learning or Control) of each participant's average classification accuracy then calculated the mean difference in accuracy between the groups. Repeating this permutation procedure 10,000 times allowed us to construct a distribution of group-level rating differences to be expected under the null hypothesis of no difference in classification accuracy between these groups. The proportion of elements in this distribution exceeding the actual mean difference score was taken as the *p* value of the difference between Learning and Control groups for a given experiment. Permutation testing was used to avoid distributional assumptions of parametric tests. Because we had a clear hypothesis about the sign of the mean difference score in each experiment (i.e. Learning > Control), *p* values were determined from the positive tail of null permutation distributions.

We also used linear models to explore the impact of reported age, reported gender, reaction time, and screen resolution on classification accuracy. These models did not reveal a significant impact of any of these factors on testing phase accuracy, and were consistent with the permutation tests in measuring performance differences between Learning and Control groups. Experiment data and analysis code are available at https://github.com/drewlinsley/co_scene.

2.2.2 Results

The human visual system automatically discovers regularities in visual information independent of task demands or conscious intent [162]. We suspected that this form of statistical learning is responsible for our earlier finding that the identities of scenes'

objects centripetally bias perception of their spatial properties [35], and would further support a bias in the opposite direction, with scenes' spatial properties biasing perception of their object features. The proposed bias reinforces the perceived identities of scenes' objects by bringing them into alignment with scenes' spatial properties. This process resolves perceptual inconsistencies in the resources and improves scene categorization accuracy by making it easier for downstream regions involved in decision-making to infer a scene category.

In Experiment 1 we investigated if observers automatically learn statistics about co-occurring objects and spatial properties in scenes, and measured how these statistics impact scene recognition accuracy. To test this, two groups of participants (Learning and Control) completed an online behavioral experiment in which both viewed sequences of novel scene exemplars in a training phase and then categorized object-ambiguous versions of those scenes in a testing phase. In the critical experimental manipulation, only the Learning group was given the opportunity to learn statistics of co-occurring objects and spatial properties during the training phase. We expected that participants' representations of these statistics would act as a perceptual look-up table, supporting rapid inference of the kinds of object features associated with these scenes' spatial properties.

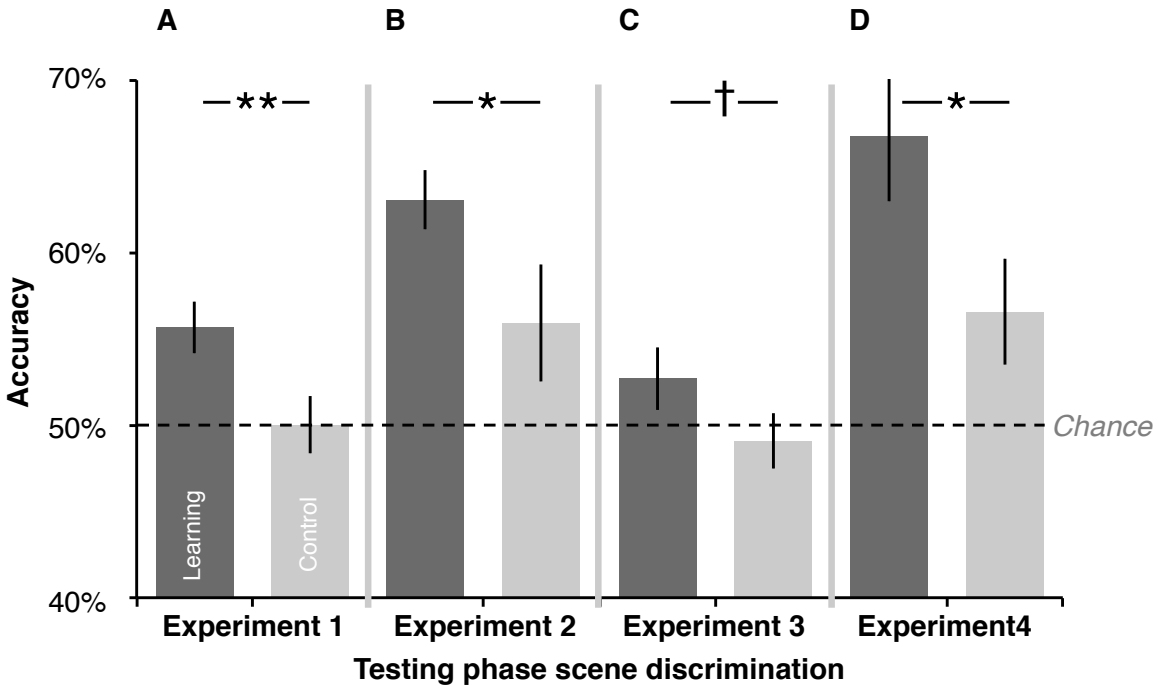


Figure 2.3. Behavioral results. (A) Scene categorization accuracy of object-ambiguous testing phase scenes was significantly better for the Learning group, trained on object-intact exemplars of Scene A and Scene B, versus the Control group, trained on object-ambiguous versions of these scenes. (B) The benefit of learning co-occurrence statistics persisted when controlling for attention between Learning and Control groups during training. (C) Additionally, co-occurrence statistic learning did not depend on sequencing exemplars with imbalanced transition probabilities. (D) Learning co-occurrence statistics for some scene exemplars generalized to others for which they were not learned, signaling parameterized storage of these statistics. Error bars are s.e.m. † : p = 0.072; * : $p < 0.05$; ** : $p < 0.01$.

2.2.2.1 Co-occurrence Statistic Learning Improves Scene Recognition

In order to understand if the Learning group successfully learned co-occurrence statistics and to measure their impact on scene recognition, we compared performance of the two groups in a subsequent testing phase. During this testing phase, both groups categorized object-ambiguous versions of the two novel scene categories. We hypothesized that if the Learning group outperformed the Control group on this testing phase task it would mean that they leveraged learned co-occurrence statistics to activate object codes consistent with scenes' spatial properties, reinforcing their categorization decisions and improving accuracy. Consistent with our expectations, the Learning group was significantly better at categorizing testing phase scenes than the Control group (55.680% versus 50.024%, d = 0.492, p = 0.007; Figure 2.3.A).

2.3 EXPERIMENT 2

It was possible that the difference between Learning and Control group performance in Experiment 1 did not reflect the influence of co-occurrence statistics but instead resulted from differences in the amount of attention paid by each group during the training phase. Since the Control group viewed object-ambiguous scenes in their training phase, it may have simply been more difficult for them to engage with the experiment during either the learning or testing phases than the Learning group. Experiment 2 attempted to replicate Experiment 1 with procedures that should have mitigated any attentional differences between the Learning and Control groups.

2.3.1 Materials and Methods

2.3.1.1 Participants

A total of 103 participants (54 males) between 19-65 years old were recruited for the experiment. Recruitment, consenting procedures, payment, and experiment duration were consistent with Experiment 1.

2.3.1.2 Stimuli

This experiment used the same stimuli as Experiment 1.

2.3.1.3 Procedure

This experiment's procedure was nearly identical to Experiment 1. The only difference was in the task performed by participants during the training phase. In contrast to the color discrimination task of Experiment 1, here participants in both Learning and Control groups performed a spatial discrimination task by judging if the currently viewed scene exemplar was more or less "spacious" than the preceding one (Figure 2.2.B). This task controlled for attention allocation during the training phase, since it required participants from both Learning and Control groups to pay attention to the spatial properties of each scene. As in Experiment 1, participants were provided feedback to encourage engagement: the currently viewed image was outlined green for a correct response or red for an incorrect response. The testing phase was identical to Experiment 1.

2.3.1.4 Experimental Setup

The experimental setup of this experiment was the same as in Experiment 1.

2.3.1.5 Data Analyses

Consistent with Experiment 1, preplanned filters were applied to this data to control for measurement error from 1) Participant disengagement, and 2) Time-of-day effects. 1) The Participant disengagement filtering procedure, which separated engaged from disengaged participants based on their performance during the training phase, excluded 35 participants from this experiment (19 Control). Note that the large number of participants filtered here relative to Experiment 1 likely reflects the difficulty of the size-discrimination task versus the color discrimination task. 2) The Time-of-day filtering procedure excluded 5 participants (2 Control) for completing the experiment at a time of day 2 standard deviations beyond each experimental group's mean. In total, 63 (29 control) of the original 103 (50 control) participants in this experiment passed these filters.

2.3.1.6 Statistical Analyses

As in Experiment 1, the number of participants in Learning and Control groups in this experiment were equalized to support statistical testing. This was done using the same bootstrapping procedure detailed for that experiment. All other statistical testing procedures were identical to Experiment 1.

2.3.2 Results

We created a new experiment to control for the possibility that the results in Experiment 1 were driven by the level engagement of Learning and Control groups throughout the experiment. This experiment controlled for engagement by having both groups perform a

spatial discrimination task during the training phase, in which participants indicated if the currently viewed exemplar was more or less "spacious" than the one preceding it. Training phase performance was similar for both groups, suggesting that it successfully controlled for attention during the training phase (Learning: 66.349% versus Control: 63.546%, $p = 0.147$).

The training phase manipulation ended up benefiting scene classification of both groups during the testing phase (Figure 2.3.B, deviation of both bars from chance). This basic result is consistent with the general impact of attention on perception: it can yield gains in discriminating visual information and drive perceptual learning [170]. Here, we suspect that attention similarly enabled both groups to more effectively encode the visual features of these scenes. For both groups, this meant becoming more sensitive to the spatial properties of the novel scene categories. The increased spatial sensitivity of this task improved performance in categorizing subsequent object-ambiguous testing scenes, which could only be discriminated on the basis of their spatial properties. However, as in Experiment 1, the Learning group was significantly better than the Control group at scene classification (63.063% versus 55.939%, $d = 0.491$, $p = 0.034$; Figure 2.3.B), indicating that co-occurrence statistic learning and not differences in attention improved scene categorization accuracy in these experiments.

2.4 EXPERIMENT 3

We were also interested in understanding whether the impact of statistical co-occurrence learning on perception of scenes' object features depended upon the design of the training

phase stimulus sequences. These sequences were structured with imbalanced transition probabilities between scene categories that promoted implicit learning of category boundaries [168,171], which we initially believed would improve the likelihood of co-occurrence statistic learning taking place. However, this design raised the possibility that scene feature co-occurrence statistic learning depended on this very structure.

2.4.1 Materials and Methods

2.4.1.1 Participants

A total of 102 participants (48 males) between 19-65 years old were recruited for the experiment. Recruitment, consenting procedures, payment, and experiment duration were consistent with Experiment 1.

2.4.1.2 Stimuli

This experiment used the same stimuli as Experiment 1.

2.4.1.3 Procedure

This experiment was created to investigate the impact of temporal regularities on learning scenes' co-occurring object and spatial property statistics. For this reason, the experiment was identical to Experiment 1 except for the order in which stimuli were sequenced during the training phase. Here, participants viewed uniformly randomized training phase sequences. The testing phase was identical to Experiment 1.

2.4.1.4 Experimental Setup

The experimental setup of this experiment was the same as in Experiment 1.

2.4.1.5 Data Analyses

Measurement error in this experiment was filtered exactly as in Experiment 1. 1) The Participant disengagement filtering procedure excluded 2 participants from this experiment (1 Control); 2) The Time-of-day filtering procedure excluded 1 participant (1 control) with for completing the experiment at a time of day 2 standard deviations outside the mean. In total, 99 (52 control) of the original 102 (54 control) participants in this experiment passed these filters.

2.4.1.6 Statistical Analyses

The same bootstrap procedure described for Experiment 1 was used here to equalize the number of participants in Learning and Control groups. This procedure supported subsequent statistical testing. All other statistical testing procedures were identical to Experiment 1.

2.4.2 Results

We tested the impact of temporal regularities on learning scenes' co-occurring object and spatial property statistics with a new set of participants who were trained on "non-informative" (i.e. uniformly random) sequences of exemplars. Although the Learning group outperformed the Control group during the testing phase (52.705% versus 49.090%, $d = 0.307$, $p = 0.072$; Figure 2.3.C), the effect size of the difference was

approximately 40% less than Experiment 1, and was only marginally significant. However, a two-way ANOVA with experiment and group membership as factors did not find a significant interaction for the role of temporal regularities in learning scene co-occurrence statistics ($F(1,194) = 0.379$, $p = 0.539$). Thus, the current data do not support the conclusion that temporal regularities play a significant role in learning scenes' co-occurring object and spatial property features.

2.5 EXPERIMENT 4

Learning scenes' co-occurring object and spatial properties presents an enormous challenge. The most obvious explanation for the object perception bias inferred from Experiments 1, 2, and 3, is that observers store associations between the object and spatial property features of every scene exemplar viewed during the training phase. Extending this information storage regime into the real world, representations of these statistics for every scene viewed throughout an observer's life would be inefficient, requiring an enormous amount of memory for storage and extensive processing to access.

An alternative approach that would ameliorate both storage and access issues is to summarize co-occurrence statistics with a set of intermediate parameters, which capture associations between commonly encountered spatial property and object features. The spatial properties of real-world scenes are not uniformly distributed across the range of physically possible spatial properties, but fall into clusters that correspond to individual categories: for example, one cluster may correspond to museum gift shops while another corresponds to galleries. Similarly, the objects observed in real-world scenes are found in

some scene categories more often than others (e.g. an oven is almost always found in a kitchen). The visual system might be able to leverage this natural clustering of scene features into an efficient storage of co-occurrence statistics by encoding links between the central tendencies of these spatial property clusters (e.g. mean spatial properties of a scene category) and object clusters (e.g. the typical single object or ensemble of objects in a scene category). In this model, co-occurrence statistics are accessed through a similar mechanism during scene categorization: scenes' spatial properties influence their perceived objects based on associations with the cluster of spatial properties most *similar* to its own (rather than its own exact values as in the aforementioned, inefficient model).

The efficient, parameter model makes the prediction that if co-occurrence statistics are "looked-up" on the basis of the cluster of spatial properties a particular scene falls within, they should generalize to scene exemplars with spatial properties that have not previously been seen. This means that Learning group participants in Experiment 1 should perceive objects in a novel scene exemplar as biased towards the values associated with its spatial properties even if they never had the chance to learn the co-occurrence statistics for that particular exemplar. We tested this theory by creating a new version of Experiment 1, in which a Learning group viewed scene exemplars that were sometimes object-visible, and other times object-ambiguous. These participants were able to generalize co-occurrence statistics learned from object-visible exemplars to object-ambiguous ones during the testing phase, resulting in significantly more accurate classification of these exemplars than the Control group.

2.5.1 Materials and Methods

2.5.1.1 Participants

A total of 81 participants (28 males) between 19-65 years old were recruited for the experiment. Recruitment, consenting procedures, payment, and experiment duration were consistent with Experiment 1.

2.5.1.2 Stimuli

This experiment used the same stimuli as Experiment 1. However, the Learning group viewed both object-visible and object-ambiguous versions of Scene A and Scene B during the training phase.

2.5.1.3 Procedure

This experiment employed a similar design as Experiment 2 to investigate if co-occurrence statistics learned for some exemplars would generalize to other similar exemplars. The training phase in this experiment consisted of Scene A and Scene B exemplars sequenced according to the graph structure in **Figure 2.2A**, some of which were object-intact while others were object-ambiguous. An object-intact exemplar was displayed whenever the sequence drew from a white or black node in the graph (1, 5, 6, or 10), whereas an object-ambiguous exemplar was displayed whenever the sequence drew from any other node (those in grayscale). Both groups completed the one-back spatial discrimination task to control for differences in engagement.

The testing phase was identical to Experiment 1. Since the control condition in this experiment is identical to Experiment 1, Control group data for this Experiment were those collected for Experiment 1.

2.5.1.4 Experimental Setup

The experimental setup of this experiment was the same as in Experiment 1.

2.5.1.5 Data Analyses

Measurement error in this experiment was filtered using the approach previously described for Experiment 1. 1) The Participant disengagement filtering procedure excluded 31 participants from this experiment (19 Control); 2) The Time-of-day filter excluded 4 participants (2 control) for completing the experiment at a time of day 2 standard deviations beyond the mean. In total, 46 (29 control) of the original 81 participants (50 control) in this experiment passed these filters.

2.5.1.6 Statistical Analyses

To support statistical testing, the number of participants in Learning and Control groups in this experiment was equalized using the bootstrapping procedure detailed for Experiment 1. All other statistical testing procedures were identical to Experiment 1.

2.5.2 Results

We tested the plausibility of efficient, parameterized co-occurrence statistic storage by investigating if associations learned for some scenes can generalize to others and bias

perception of their objects. We tested for this effect with a modified version of Experiment 2. During the training phase, the Learning group viewed sequences of Scene A and Scene B exemplars that systematically varied between object-visible and object-ambiguous. This paradigm allowed us to measure if participants generalized co-occurrence statistics learned for object-intact exemplars to object-ambiguous exemplars. As in Experiment 2, attention and engagement were controlled with a spatial discrimination task during the training phase (no significant difference between groups: Learning: 60.475%; Control: 63.546%, $p = 0.142$).

We compared Learning and Control groups in categorizing testing phase scene exemplars drawn from grayscale nodes in **Figure 2.2.A** (numbered: 2, 3, 4 in Scene A & 7, 8, 9 in Scene B). We found that the Learning group was significantly better than the Control group at categorizing testing phase scene exemplars drawn from these nodes, for which they did not have the opportunity to learn co-occurrence statistics (66.751% versus 56.570%, $d = 0.754$, $p = 0.020$; Figure 2.3.D). Note that replicating this experiment with the training task from Experiment 1 (color discrimination) instead of the spatial discrimination task used here revealed a similar generalizability of co-occurrence statistics, albeit with an approximately 50% reduction in effect size (Appendix B Figure 1).

2.5.3 Experiment 4 Discussion

The visual system leverages information about co-occurring object and spatial property information during scene categorization to reinforce the perception of each of these resources. While statistics could be encoded for every newly encountered scene, a more

efficient regime is to maintain a set of parameters that summarize the kinds of objects and spatial properties that typically co-occur across *many* scenes. This would both reduce the amount of storage needed to maintain these statistics and, in the process, make it easier to access them during perception.

Here we find evidence for parameterized storage of object and spatial property co-occurrence statistics. During the training phase, Learning group participants viewed object-intact exemplars only when their sequence drew from specific nodes (1, 5, 9, or 10), and object-ambiguous exemplars the rest of the time. That the Learning group still outperformed the Control group at categorizing testing phase exemplars – specifically those for which *neither* group learned co-occurrence statistics – reveals that they generalized from co-occurrence statistics for other, similar scenes. Generalization is not possible from a storage regime in which co-occurrence statistics are tracked for every encountered scene, and therefore favors parameterized storage.

2.6 GENERAL DISCUSSION

Our experiments demonstrate that humans automatically learn statistics about the co-occurring object and spatial property features in computer generated versions of real-world scenes. When given the opportunity to learn these statistics, participants' performance on a scene categorization task reliably improved in all four experiments. Although statistical learning occurs across many perceptual tasks [161,162,172,173], the work here is to our knowledge the first indicating that observers capture associations in

visual features spanning local (i.e. object) and global (i.e. spatial property) scales of a scene.

2.6.1 Object and Spatial Property Combination Through Statistical Learning

But exactly how does co-occurrence statistical learning improve scene recognition? That observers were *more* accurate at categorizing scenes based on their spatial properties following co-occurrence statistic learning indicates that these statistics filled-in object information in these scenes during perception. This provided downstream regions responsible for categorization decisions (e.g. prefrontal cortex) with object features in addition to the spatial property information that was evident in scenes even if co-occurrence statistics were not learned. Although previous research has indicated that statistical learning can both yield more efficient attention allocation [74,174–177] and exploit temporal regularities [166,168], we found similar benefits for scene categorization induced by statistical learning when controlling for each of these factors (Experiments 2 and 3, respectively).

Our findings of an internal representation of co-occurring object and spatial layout features provides a mechanistic explanation for the finding outlined in Chapter 1, in which scenes' encoded spatial properties were biased by category-informative objects towards values associated with those objects [35]. This effect mirrors the object feature bias found in each of the experiments described here, and suggests that it results from a similar process dependent on co-occurrence statistics, which reinforces the perceived spatial properties in scenes with information about their objects. When considering both studies together, it appears that the visual system's access to statistics describing co-

occurring object and spatial property features supports a bias that can impact either resource (i.e. from objects to spatial properties and vice versa).

2.6.2 Statistical Learning in Scene Recognition

Statistical learning took place during an initial training phase in each experiment, where participants viewed sequences of scene images while completing either a color- or spaciousness-discrimination task. Although participants may have automatically engaged in scene recognition while viewing each scene, these tasks ensured that statistical learning was independent of task demands, falling into the category of unsupervised learning [178–180]. This learning also did not depend on either paying specific attention to the objects in scenes (as demonstrated by the attention task in Experiment 2) or temporal regularities in stimulus sequences (as demonstrated by the sequence construction in Experiment 3), each of which can benefit performance on later recognition tasks [74,174–176]. These findings indicate that the co-occurrence statistical learning observed here may be driven by neural systems that are to some extent distinct from previous statistical learning accounts [168,181].

We also found that co-occurrence statistics learned for certain scene exemplars were generalized to other, similar exemplars during later scene categorization (Experiment 4). This result aligns with prior accounts of efficient statistical learning [166,173,182], and motivates a model for how observers store and interact with these co-occurrence statistics. Observers interface with stored object and spatial property co-occurrence statistics using either an INSERT or QUERY operation [183]. One possible scenario for this interface is if a scene's objects and spatial properties are easy to

perceive, an observer uses the INSERT operation to update the appropriate parameters in the table with the new information. Alternatively if the objects in a scene are obscured as in this experiment, the observer can use a QUERY operation to look up the identities of objects that typically co-occur with the scene's spatial properties. While the INSERT operation would likely recruit memory systems and top-down control to consolidate information, the QUERY operation could occur rapidly during bottom-up recognition. Additional work is needed to identify the neural foundation for this learning and continue to explore its role in dominant feedforward models for scene recognition [67,73,184,185].

One question not addressed by these experiments is the time course of learning scene feature co-occurrence statistics: would more training yield even better testing phase performance than observed in these experiments? Given that participants learned these statistics from significantly fewer scene exemplars (training phase consisted of a combined 154 exemplars) than is typical in their daily lives, improved performance with more training is expected. The web-based nature of our experiments makes it difficult to directly address this question, as longer participation time would have likely also increased measurement noise. Nevertheless, exploring the observed bias of object perception as a function of the amount of statistical learning could provide valuable insight into this question.

2.6.3 Conclusion

We found evidence for an influence of statistical learning on scene categorization. Observers automatically learn the co-occurrence of objects and spatial properties in scenes and leverage these statistics to improve recognition, particularly when viewing a

scene with features that are difficult to discern. A key question not addressed by the current research is exactly what kinds of visual features are captured by these co-occurrence statistics. It is unlikely that these features correspond to summary statistics of scene information [186] (e.g. the average color, line orientation, or object in a scene [187,188]), and more work is needed to develop and validate alternatives.

3.0 VENTRAL VISUAL CORTEX LEARNS OBJECT AND SPATIAL PROPERTY CO-OCCURRENCE STATISTICS DURING SCENE CATEGORIZATION

Manuscript *in preparation*

During scene categorization, the visual system draws upon statistics describing co-occurring object and spatial property features to reinforce the perception of each resource. In Chapter 1, this effect resulted in a systematic bias in the perception of scenes' spatial layout, whereas the reverse was found in Chapter 2: scenes' spatial layout was apparently leveraged to "fill-in" in the identities of perceptually masked objects. Despite this significant role of object and spatial property co-occurrence statistics in scene categorization, the cortical regions responsible for learning them are unknown. While intuition would suggest that cortical regions functionally associated with learning (e.g. medial temporal lobe) would also capture these statistics, recent research into visual statistic learning suggests that the visual system may also have this capacity. Here we provide evidence that regions of ventral temporal cortex (VTC) implicated in perceptually processing scenes' objects and spatial properties are also involved in learning their co-occurrence. The activity profiles of these regions are driven by the amount of object and spatial property co-occurrences in scenes. A functional connectivity analysis identified participation of other regions in VTC strongly associated with processing of scene information. These results indicate a dual role for VTC in scene recognition: in both processing visual features in scenes and learning their statistical regularities.

3.1 INTRODUCTION

In order to deal with the vast amount of information in a visual scene, humans leverage knowledge about commonly repeating elements. For instance, straight lines are a useful

cue for navigating through the world as they are more commonly present in man-made than natural environments [15,156]; likewise, an object moving through a visual scene (such as a car) can be decoded by identifying temporally coherent low-level features such as edges [189]. A large body of research indicates that humans capture these regularities during passive viewing, automatically learning the statistical structure of visual information without behavioral supervision. Indeed, statistical learning plays a core role in human development, with evidence of this benefit first appearing in early infancy [180,190,191] and imparting a lasting impact throughout life by improving information processing across sensory modalities [161,176,192,193].

However, little is known about the role of statistical learning in many core visual processes. Scene categorization is one such function that is crucial for human behavior, supporting effective behavior and navigation through the world. It is dependent on fast and accurate perception of two complementary resources of information within a scene. One of these resources is the identities and spatial relationships of objects in an environment, which are often sufficient to recognize a scene [1,2,153,194]. In contrast, more recent research has demonstrated that the spatial properties of a scene, such as its size, form a rich informational resource during rapid scene categorization [15,116].

We found evidence that statistical learning extends to scene categorization by influencing how scenes' object and spatial property features are processed. In Chapter 1, we demonstrated that scenes' perceived spatial properties are biased towards the values associated with their objects – causing a room with an oven to be perceived as more spatially similar to the average-sized kitchen than if it did not have an object associated with kitchens [35]. Evidence for the opposite bias was described in Chapter 2, in which

implicitly learned object and spatial property co-occurrence statistics were leveraged during scene categorization. When presented with scenes containing perceptually masked objects, participants who had learned these statistics "filled in" these objects with features associated with their spatial properties. These complementary perceptual biases reveal the significant impact of scene object and spatial property co-occurrence statistics on scene categorization: the associations reinforce perception of each resource, which reduces perceptual conflicts between scene categories and improves accuracy.

How are these statistics stored in the brain? Although there is extensive work describing the neural foundations for learning statistics describing the temporal or spatial context of objects [168,171,195,196] in support of object recognition, it's unclear if those same mechanisms also capture the co-occurrence statistics of object and spatial property features in a way that supports scene categorization. Here we took the first step in this exploration by asking what regions of the brain have activity profiles that indicate participation in this form of statistical learning.

We used functional magnetic resonance imaging (fMRI) to record brain activity while participants viewed exemplars from novel computer-generated scene categories. As the main experimental manipulation, each scene category varied in their number of potential object and spatial property associations. This manipulation was instantiated by altering the visibility of objects (i.e. intact versus ambiguous) and range of spatial properties occupied by each scene category (i.e. spanning a wide versus narrow range of sizes). Exemplars from each category were sequenced in a block design to control for perceptual effects driven solely by either scene spatial properties or objects. We expected that this design would identify regions of cortex with activity profiles modulated by the

amount of co-occurring object and spatial property features within each scene category, which we considered evidence of involvement in this form of statistical learning.

Although prior work has investigated neural systems recruited when drawing upon object and spatial context associations [reviewed in 197], the current study is distinct in several ways. Those studies explored neural activity elicited when explicitly recalling or otherwise leveraging previously learned, *semantic level* associations between a single object and the spatial context in which it is typically found [167,198,199]. In contrast, the current experiment measures neural activity elicited while learning associations in novel scene categories. Since participants were not told the names of these categories, and the experimental task did not require semantic-level processing, we reasoned that this learning activity must be related to forming *perceptual level* associations between scenes' object and spatial property features.

Given evidence of the involvement of ventral temporal cortex (VTC) in processing visual features for scene recognition [7,12,19,24,35,116,200] as well as visual statistic learning [171,195], we hypothesized that it would similarly participate in the co-occurrence statistical learning tested here. We also expected to find through an analysis of functional connectivity that participating regions would systematically engage with other cortex throughout the experiment, potentially for processing scene features or ultimately storing co-occurrence statistics.

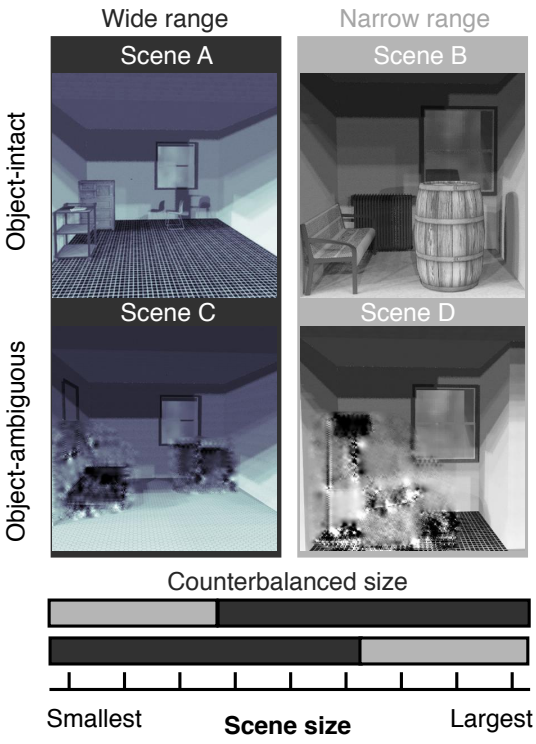


Figure 3.1. Experiment stimuli and procedure. Computer generated versions of novel scenes were created, which were defined by object and spatial property combinations that are unlikely in the real world. Scenes were either object-visible or object-ambiguous and occupied a wide or narrow size range (representing spatial properties). Object-visible categories were Scene A (chair, cart, and dresser) and Scene B (bench, radiator, and barrel); object-masked categories were Scene C and Scene D. In contrast, Scene A and Scene C were generated with a Wide size range, while Scene B and Scene D were generated with a narrow size range. The precise spatial scale of each scene category was counterbalanced across participants ("Counterbalanced size"; gray bars corresponds to the Narrow range and black to the Wide range). This meant that for half of the participants (e.g.) Scene B was larger than Scene A despite its exemplars occupying a narrower range of sizes. All scene exemplars were randomly presented in either blue or gray scale.

3.2 MATERIALS AND METHODS

3.2.1 Participants

Fourteen participants (9 female, aged 18 – 29 years) gave written informed consent in compliance with Boston College Institutional Review Board procedures approved procedures. These participants satisfied typical selection criteria and were right handed, of normal or corrected-to-normal visual acuity, and had no history of neurological disease. Two participants with excessive motion artifacts were excluded from analysis. Participants were paid $60.

3.2.2 Stimuli

Visual stimuli were computer-generated exemplars of two "novel" scene categories ("Scene A" and "Scene B", Figure 3.1), which contained a combination of objects and scene spatial properties that are not reliably observed together in the real world. Novel rather than familiar scene categories were used to directly manipulate how participants learned object and spatial property co-occurrence statistics.

These categories were created by placing distinct ensembles of objects in indoor rooms drawn from two adjacent and non-overlapping ranges of floor area (which established different spatial properties for the two categories). The object ensembles assigned to both scene categories occupied similar spatial extents within the rooms. Scene A contained a chair, cart, and dresser, whereas Scene B contained a bench, radiator, and barrel. Object models in each scene image were randomly drawn from a

pool of 10 for each type of object, and positioned in one of 24 randomly selected configurations.

Both scene categories contained 150 unique exemplars, with floor areas ranging in 0.035 square meter increments from 4.5 square meters – 15 square meters. Scene A always occupied a wider range of floor areas ("Wide range") than Scene B ("Narrow range"; Figure 3.1), and therefore had more potential object and spatial property associations because its object ensemble linked to more values of floor area. As detailed below in Procedure, the precise spatial range of the scene categories (i.e. if Scene A was larger than Scene B or vice versa) was counterbalanced across participants to control for veridical perceptual effects, which were not of interest in this study. Rooms were rendered using Trimble Sketchup (www.sketchup.com), IRender nXt 4.0 (www.renderplus.com), and custom Ruby scripts. Separate versions of each image were produced in blue and gray scale and cropped to 400 x 400 pixels.

In addition to the "object-intact" categories Scene A and Scene B, we produced "object-masked" versions of each: Scene C and Scene D. These categories were spatially identical to either Scene A (Scene C; "Wide range") or Scene B (Scene D; "Narrow range"), but contained ineffable wavelet-scrambled [113] versions of a couch, lamp, and cabinet (i.e., objects not associated with either new category; "Object-ambiguous", Figure 3.1). The use of perceptually masked objects was preferable to excluding objects from these rooms to maximize their spatial similarity to object-intact exemplars in Scene A and Scene B.

3.2.3 Procedure

Human observers automatically learn statistics describing scenes' co-occurring object and spatial property features, which can bias perception of these resources and ultimately improve scene recognition. The general strategy of this experiment was to record functional brain volumes while participants were given the opportunity to engage in this automatic learning process in order to understand where it takes place in the brain. Although some forms of visual statistical learning can take place rapidly, needing only a few trials [201], we wanted to measure neural processes related to learning scene feature statistics that spanned entire scene categories. This required participants to view a large amount of exemplars, particularly for Scene A and Scene B. To ensure that we could easily identify the statistical learning activity elicited by each scene category we adopted a block-design to measure this activity.

Participants completed a block-design fMRI experiment consisting of 8 scan runs, in which each scan run contained images of scene exemplars from one of the four novel scene categories described above (8 total stimulus blocks total per participant; this design is inspired by others devoting extensive scanning time to measure activity during stimulus encoding [202,203]). We planned to randomize the order of the 4 stimulus blocks (each with a different scene category) within each half of the experiment (consisting of 4 scan runs), but a technical issue meant that all but one of the participants viewed the same block order: run one contained images from Scene C, run two scene B, run three Scene A, run four scene D, run five scene C, run six scene A, run seven scene D, run eight scene B. We do not believe that this affected our findings because it is unlikely that this specific block order could cause the pattern of activity discussed in the

Results. In other words, our choice for block order randomization was not correctly implemented, but was also an overly conservative design choice that did not impact the outcome of the experiment.

To control for perceptual effects related to scenes' veridical spatial properties (i.e. if Scene A was always larger than Scene B), the spatial properties occupied by each scene category was counterbalanced across participants (Figure 3.1, "Counterbalanced size"). For half of the participants, exemplars from Scene A and Scene C were larger than Scene B and Scene D, while the opposite was true for the other half. While we could have tried to control for these perceptual effects of non-interest by regressing them out of estimates of neural activity, controlling them through experimental design is more reliable [204].

Scene exemplars were shown for 1500ms, during which time participants were instructed to use a button box to judge whether the preceding scene was more or less "spacious" than the current scene. This task was chosen because it ensured similar levels of participant engagement when viewing object-intact versus object-ambiguous scenes (see Chapter 2.3). Scenes were also outlined in green following correct responses and outlined in red following incorrect responses to further engage participants in the task.

Each scan run contained 252 exemplars drawn uniformly random from a single scene category and lasted 6 minutes 48 seconds, including 30 seconds of fixation at the end. In other words, each scan run contained a stimulus block of exemplars from a single category, which we expected would maximize our ability to capture the complete time course of statistical learning evoked by the category. Experimental stimuli occupied the central ~ 4.4˚ of visual space. Both stimulus presentation and behavioral data collection

were executed with custom MATLAB code using the Psychophysics toolbox[114], which can be found at https://github.com/drewlinsley/co_scene.

Scan sessions also included two functional localizer scans, each of which lasted 7 minutes 33 seconds. During these scans participants viewed blocks of color images of scenes, faces, objects, and scrambled objects, presented at a rate of 1.33 images per second and occupying the central ~ 10.25˚ of visual space [119].

3.2.4 MRI Acquisition

All scan sessions were conducted at the Brown University MRI Research facility using a 3T Siemens PrismaFit scanner with a 64-channel head coil. Structural T1* weighted images for anatomical localization were acquired with 3D MPRAGE pulse sequences (TR = 1900 ms, TE = 3.02 ms, TI = 950 ms, voxel size = 1 x 1 x 1mm, matrix size = 256 x 256 x 160). T2* weighted scans sensitive to blood oxygenation level-dependent (BOLD) contrasts were acquired with a gradient-echo echo-planar pulse sequence (TR = 3000ms, TE = 25ms, voxel size = 3 x 3 x 3 mm, matrix size = 64 x 64 x 45). Images were rear projected onto a screen at the head end of the scanner bore and viewed with a mirror attached to the head coil. The projected field subtended 21˚ x 13˚ at 1920 x 1200 pixel resolution.

3.2.5 fMRI Data Analysis

Standard preprocessing routines were applied to functional volumes, including resampling volume slices in time to match the first slice of each volume, spatially

realigning scan volumes to the first volume of each scan, and spatial normalization to the Montreal Neurological Institute (MNI) template. Preprocessed scan volumes were also spatially smoothed with an 8 mm FWHM Gaussian filter.

Activity in each voxel was analyzed with a general linear model (GLM) implemented in SPM12. These models produced beta volumes for each scan run, capturing hemodynamic response function (HRF) convolved activity elicited by each scene category (convolution supported identification of activity related to scene stimuli, excluding undesired contributions such as from the fixation-cross or the behavioral task). Each GLM included an autoregressive AR(1) model to account for serial noise, filters that removed low frequency signal drifts, and nuisance regressors to account for global signal variations, participant motion, and reaction times for the orthogonal spatial discrimination task.

We analyzed involvement in scene feature co-occurrence statistic learning with first level (i.e. for each participant) linear contrasts that identified voxels with greater differences in activity between [Scene A and Scene B] than [Scene C and Scene D]. We expected differences in activity between the scene categories since each was generated with a different combination of object information (object-intact versus object-ambiguous) and range of sizes (wide versus narrow). However, we were specifically interested in identifying activity related to learning the object and spatial property co-occurrence statistics within each category. Statistical learning activity was captured with a contrast measuring an activity interaction across these categories: [Scene A * (1), Scene B * (-1), Scene C * (-1), Scene D * (1)]. We considered voxels to be involved in scene feature co-occurrence statistic learning if their mean contrast value at a second level (i.e.

across participants) was significantly greater than 0. We did not test for main effects of objects (i.e. visible or masked) or spatial properties (i.e. wide or narrow), as these were not relevant for identifying voxels participating in statistical learning.

3.2.6 gPPI Analysis

We were also interested in identifying regions with trial-wise contributions to co-occurrence statistic learning that may not have been evident through our main GLM analysis. To do this, we estimated task-dependent functional connectivity between suprathreshold voxel clusters from the second level of the main GLM analysis (seed regions) and every other cortical voxel in each participant's brain.

Functional connectivity was estimated with the generalized psychophysiological interaction toolbox (gPPI) [205] implemented in SPM8. Linear models were fit at each voxel that simultaneously estimated the main effects of task activity (i.e. related to the onset of each scene category) and both task-dependent and -independent correlations between that voxel's HRF convolved time course and the average convolved time course in each seed region.

We established functional connectivity by calculating first level activity interactions across scene categories (as defined above) with task-dependent connectivity beta volumes. In other words, this identified voxels with functional connectivity to a seed region that was modulated by the amount of co-occurrence statistics in a scene category. Voxels with mean second level contrast values significantly greater than 0 were considered as being functionally connected to the appropriate seed region.

3.2.7 Statistical Analysis

Participant contrast volumes from each GLM and gPPI analysis were statistically thresholded at a second level with exact permutation tests. In each case, the deviation of contrasts from 0 across subjects was converted to *t* values. These *t* values were calculated with linear regressions that contained nuisance regressors controlling for participant age, gender, and the scene categories' veridical spatial property ranges (i.e. if Scene A was larger than Scene B or vice versa). Voxels were thresholded at $p < 0.001$ based on the positive tail of null distributions calculated across the full set of 2^{12} possible sign permutations in these volumes [123]. Suprathreshold voxel clusters in the volume of observed *t* values were considered significant if their size was exceeded by fewer than 5% of elements in the distribution of maximum cluster sizes gathered from null volumes thresholded at $p < 0.001$. Cluster extent sizes were 64 voxels for the GLM analysis, 60 voxels for the left visual seed gPPI, 59 voxels for the right visual seed gPPI, and 79 voxels for the right frontal visual seed gPPI.

3.2.8 Regions of Interest

Regions of interest (ROI) were defined through a two-step procedure applied to data from localizer scans that made ROI definition more reliable across participants. First, in a cross-validation procedure, each contrast of interest (e.g. scenes > objects) was thresholded at $t = 2$ for all but one subject. Volumes of the thresholded subjects were summed together and further thresholded to produce a group volume only containing voxels active for more than 50% of subjects. The left out subject's contrast of interest

was then thresholded at $t = 2$ and masked with this group-level volume. The procedure was repeated for every participant so that ROI definitions for each were based on activation maps from all other participants.

The procedure was performed for the contrast of scenes > objects (to identify parahippocampal place area (PPA), retrosplenial complex (RSC), and transverse occipital sulcus (TOS)), objects > scrambled objects (lateral occipital (LO) and posterior fusiform sulcus (pFs) subdivisions of lateral occipital complex (LOC), and scrambled objects > objects (early visual cortex (EVC)). ROIs were labeled by hand in contrasts that defined multiple ROIs.

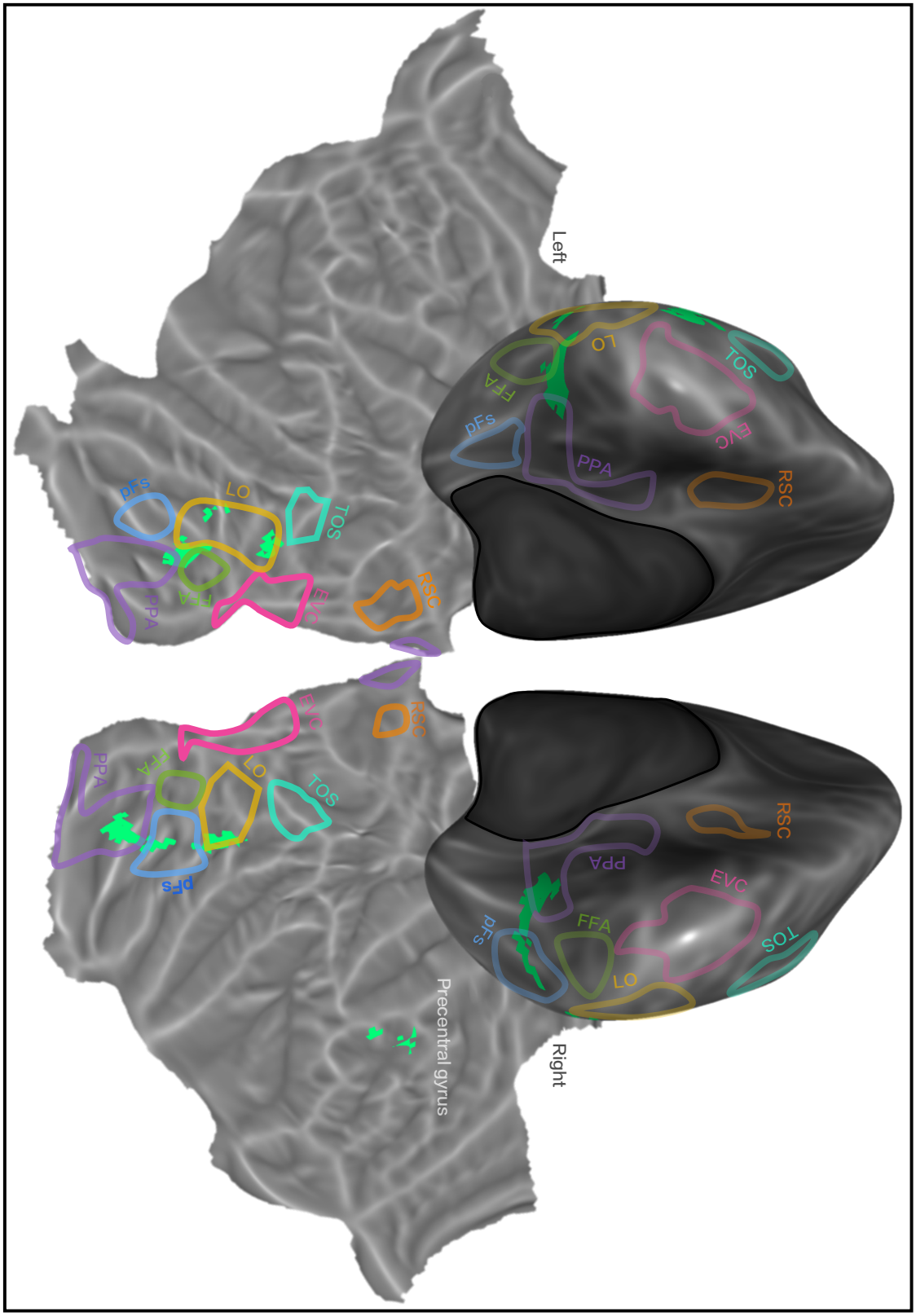
In order to label suprathreshold voxel clusters from the GLM and gPPI analyses, we calculated the percent voxel overlap with each ROI. A voxel cluster was assigned the label of an ROI if these percentages were significantly greater than 0 across subjects following Bonferroni correction. Clusters falling outside the bounds of any functionally defined ROI were anatomically labeled with the Neuromorphometrics atlas in SPM12.

3.3 RESULTS

As demonstrated in Chapter 2, scene categorization is improved when observers are first given the opportunity to learn regularities of co-occurring object and spatial property features in scene categories. This improvement is evidence that these statistics enforce systematic biases in the perception of each other: scenes' objects can bias perception of their spatial properties (Chapter 1 [35]) and vice versa (Chapter 2). Here we used fMRI to locate regions of cortex involved in learning the statistics that support these biases.

Participants viewed images from novel scene categories sequenced in a block design while their brain activity was measured. Each block displayed exemplars from one of four computer generated and novel scene categories. These scene categories had different amounts of object information visible to participants and occupied distinct ranges of spatial properties, making each unique in its total amount of potential co-occurrence statistics to be learned. After controlling for effects driven solely by differences in the perceptual properties of the scene categories, we isolated activity related to learning these co-occurrence statistics.

Figure 3.2. GLM results. First level GLM analyses identified voxels with significant Object X Size interactions. Voxels with mean responses significantly greater than 0 at the second level were considered involved in learning scenes' object and spatial property co-occurrence statistics. Permutation tests thresholded voxels at $p < 0.001$ and clusters at $p < 0.05$. Suprathreshold voxels are plotted in bright green; all other colors correspond to ROI definitions.


Left
LO
TOS
FFA
pFs
EVC
PPA
RSC
pFs
LO
TOS
FFA
PPA
EVC
RSC
EVC
RSC
PPA
FFA
LO
TOS
pFs
RSC
PPA
EVC
pFs
FFA
TOS
LO
Precentral gyrus
Right

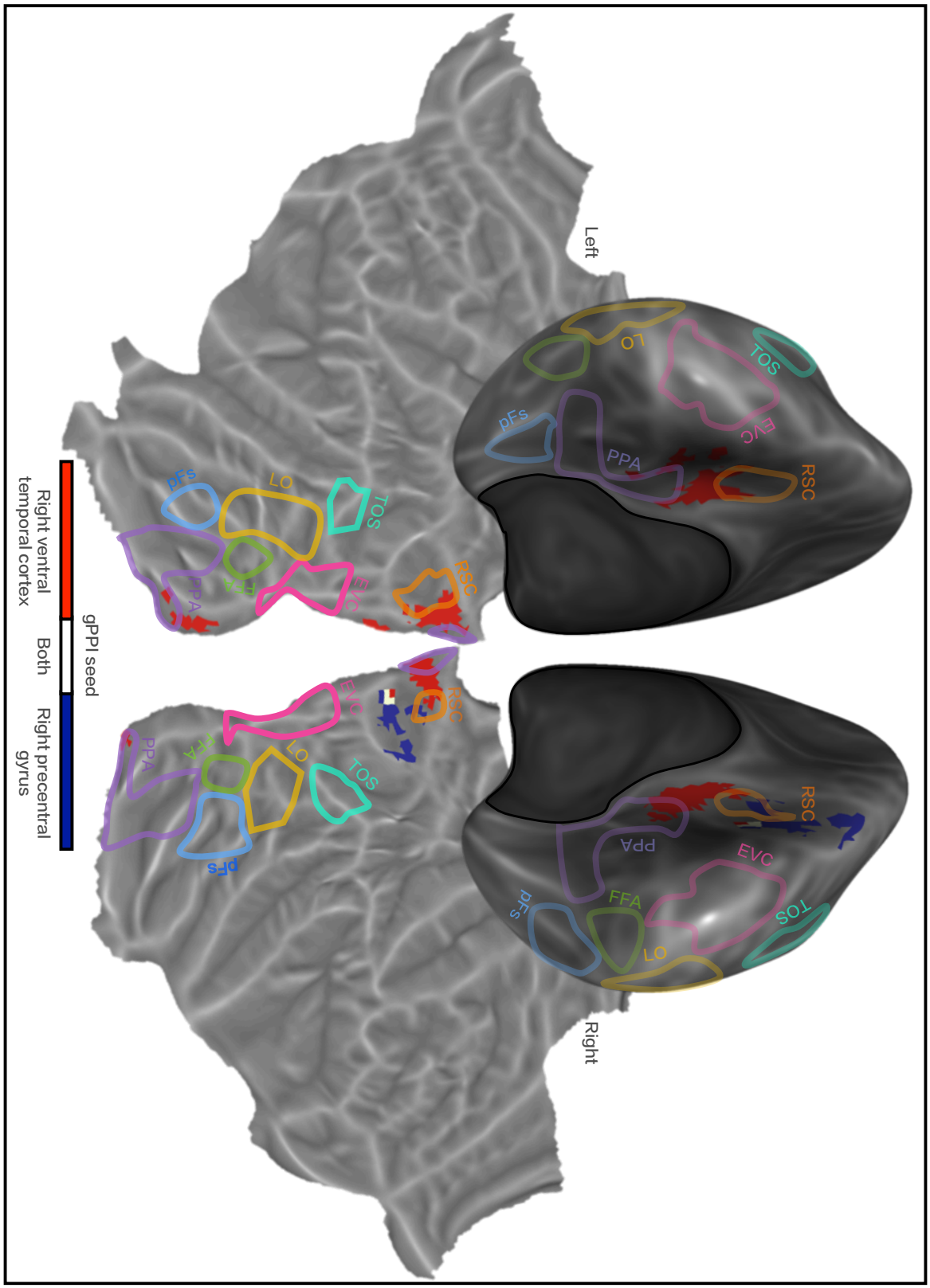
3.3.1 GLM Analysis

Participants viewed images of scene exemplars belonging to scene categories that were crossed in the amount of object and spatial property information they contained. Exemplars in Scene A and Scene C spanned wide spatial ranges whereas Scene B and Scene D spanned narrow spatial ranges. In contrast, Scene A and Scene B were object-intact while Scene C and Scene D were object-ambiguous. We expected that voxels involved in object and spatial property co-occurrence statistic learning would have activity profiles driven by the number of unique co-occurrences in a scene category. In other words, a region involved in co-occurrence learning would respond more strongly to Scene A blocks (object-intact and wide spatial range) than Scene B blocks (object-intact and narrow spatial range). However, it's possible that this same response profile could have been identified regions that were sensitive to the range of spatial properties in categories, rather than the number of object/spatial property co-occurrences to be stored. We controlled for this factor by identifying voxels for which the difference in activity elicited by Scene A versus Scene B was significantly larger than the difference between object-ambiguous versions of these categories (Scene C versus Scene D). Since co-occurrence statistics could not be learned for these object-ambiguous scene categories, this contrast constrained our analysis to identify voxels driven by co-occurrence statistics of the scene categories.

Consistent with our predictions, two clusters of suprathreshold voxels were located in VTC (Figure 3.2). One cluster overlapped with participants' left PPA and left

LO (99 voxels centered at [-30, -76, -16]) while the other cluster overlapped with right PPA (142 voxels centered at [42, -61, -17]). An additional cluster of voxels was revealed in right frontal cortex. Because it fell outside the bounds of any functionally localized ROI, it was assigned its anatomical label of right Precentral Gyrus (PrG; 65 voxels centered at [45, 11, 26]).

Figure 3.3. gPPI results. gPPI analyses measured task-dependent functional connectivity between regions involved in learning scene feature co-occurrence statistics and every other voxel in the brain. Red voxels displayed suprathreshold connectivity with right ventral temporal cortex; blue voxels with right precentral gyrus; all other colors correspond to ROI defintions. Permutation tests thresholded voxels at $p < 0.001$ and clusters at $p < 0.05$.


Left
Right
LO
TOS
pFs
EVC
PPA
RSC
FFA
gPPI seed
Right ventral temporal cortex
Both
Right precentral gyrus

3.3.2 gPPI Analysis

Given evidence of VTC and PrG participation in object and spatial property co-occurrence statistic learning, we were next interested in expanding our search to identify other cortical regions contributing to this learning that may not have been evident in the initial GLM search. We expected that contributions to co-occurrence learning, such as encoding scenes' features or storing the statistics, might be evident through changes in functional connectivity with VTC or PrG that was modulated by the amount of co-occurrence statistics in scene categories. Voxels functionally connected to VTC and PrG were identified with gPPI, which estimated task-dependent functional connectivity between these regions and every other voxel in the brain (Figure 3.3). Right VTC was connected with a cluster overlapping right PPA and right RSC (106 voxels centered at [24, -52, -4]) and an additional cluster overlapping left RSC (138 voxels centered at [12, -61, 2]). Note that the cluster of voxels overlapping with participants' right PPA was anterior to the right VTC cluster from the main GLM analysis and did not overlap. Right PrG was connected with a cluster overlapping right RSC (109 voxels centered at [18, -67, 14]). Right VTC and right PrG functional connections with RSC were mostly independent, with an overlap of 4 voxels.

3.4 DISCUSSION

Human observers automatically learn the statistical structure of the visual environment, capturing regularities across both space and time [161,162,166,171,195]. We recently

found that this statistical learning extends to the domain of scene recognition. In Chapter 2, we found that participants were significantly better at categorizing scene exemplars with perceptually masked objects when given the opportunity to learn statistics about the scenes' co-occurring object and spatial property features beforehand. This indicated that object features associated with the scenes' spatial spatial properties "filled-in" their perceptually masked objects, joining the similar and complementary study detailed in Chapter 1 [35] in suggesting that the visual system utilizes co-occurrence statistics to reinforce perception of scene features. Here, we used fMRI to identify regions of cortex involved in the process of learning these co-occurrence statistics

3.4.1 Statistical Learning of Co-occurrence statistics

Our main finding is that overall activity in right PrG and bilateral VTC corresponded to the quantity of co-occurrence statistics in a set of novel scene categories. The sensitivity of PrG to these co-occurrence statistics aligns with prior evidence for its involvement in statistical learning, in which it was shown to have a response profile linked to stimulus familiarity and temporal prediction [195,196,206]. It is therefore possible that our finding of PrG involvement indicates that learning co-occurrence statistics of scenes' object and spatial property features utilizes similar neural circuits involved in capturing regularities in temporal structure and spatial context.

On the other hand, the involvement of VTC suggests that overlapping populations of neurons may participate in both encoding scenes' object and spatial property features and also learning their co-occurrences. Within VTC, voxels sensitive to co-occurrence statistics overlapped with bilateral PPA and LO, each of which has previously been

implicated in scene recognition [12]. PPA is associated with processing multiple dimensions of scene features: it activates more strongly to images of scenes than objects or other types of visual stimuli [24], encodes scene information correlating with both spatial properties [19,35,207] and scene categorization decisions [12], and is sensitive to the identities and locations of objects in scenes [35,39,208]. In contrast, LO has an activity profile linked with scenes' objects, and it encodes information about the identities and configurations of objects in scenes [7,30]. However, a growing body of research has also associated both of these regions with visual statistic learning. Each encodes information about the temporal context of stimuli and responds more strongly to stimuli with a familiar temporal structure or spatial context [171,195,209].

The current results therefore serve two key purposes. First, they extend the role of PPA and LO in statistical learning to indicate an additional involvement in capturing the statistical co-occurrence of object and spatial property features. This is the first demonstration of visual statistic learning in which associations combine features that describe different scales of visual space (i.e. linking a scenes' local object features with its global spatial property features). Second, these results make the prediction that structures involved in processing visual features may be either proximal to or interdigitated with those involved in statistical learning. Future research is needed to explore these structures in a higher resolution than the current experiment in order to disambiguate their computational contributions to perception and statistical learning.

The current results do not define the actual role of VTC and PrC in statistical learning, for which two routes for participation can be inferred from previous research. One possibility is that these regions may contribute to the process of storing co-

occurrence statistics. This account is based on behavioral evidence showing that these statistics are stored using a limited set of "parameters", which summarize co-occurrence statistics across multiple scene exemplars and are updated – rather than added to – with experience (Chapter 2). For instance, one such parameter capturing the association of televisions with "den-sized" rooms might be updated to reflect visual features of contemporary televisions. Within this framework the current results may reflect VTC and PrC involvement in updating these parameters. Another possibility is that VTC and PrC activity profiles reflect their utilization of co-occurrence statistics during scene recognition: these regions are using co-occurrence statistics to reinforce perception of a scene feature. This alternate account is far more plausible for VTC than PrC, given its involvement in perception of scenes' objects and spatial properties.

3.4.2 Functional Connectivity

We were also interested in understanding what cortical areas VTC and PrC recruited during co-occurrence statistic learning. gPPI identified clusters in left PPA and bilateral RSC as having functional connections to VTC and PrC that were modulated by the amount of co-occurrence statistics in each scene category. In parallel with past research on the PPA response profile, RSC preferentially responds to scenes over other stimuli, and both captures and retrieves information about the environment that is crucial for navigation [42,210,211]. Together, these regions have been considered as crucial nodes in a scene-processing network that has access to regions of medial temporal and prefrontal cortex involved in memory-based processing [212].

Nevertheless, the current research cannot clarify the role of these regions in statistical learning of object and spatial property co-occurrence statistics. One possibility is that their functional connectivity reflects processing of visual features that are later utilized by VTC and PrC for co-occurrence statistic formation. Another possibility is that they are actually involved in long-term storage of parameters describing these statistics, either carrying this process out themselves or through interactions with other connected regions involved in memory formation.

3.4.3 Conclusion

Here we have described an initial exploration of the neural foundations of co-occurrence statistical learning during scene recognition. Bilateral VTC and right PrG participated in this process, with each possessing activity profiles related to the amount of co-occurrence statistics in a set of novel scene categories. These regions further demonstrated functional connectivity with regions of VTC involved in processing scene features. Our results reveal a network of regions located in both visual and frontal cortex that is active while statistics describing co-occurring object and spatial property features in scenes are learned. As these results only serve to identify regions involved in this process, future research must investigate the computational contributions of each of these areas during statistical learning. A greater understanding of these neural computations will prompt revisions of dominant feedforward models for scene recognition to account for the unsupervised statistical learning reported here [67,185].

4.0 OBJECT AND SPATIAL PROPERTY CROSSTALK IMPROVES SCENE RECOGNITION

Manuscript *in preparation*

Contrasting with the dominant feedforward model for scene categorization, the framework presented in the General Introduction holds that humans categorize scenes based on information about scenes' object contents and spatial properties that becomes entwined during the encoding stage. This leads to systematic biases in perception of both resources of information, as discussed in Chapters 1 and 2. This Chapter outlines a modeling approach to answering the question of whether encoding-stage biasing of scenes' local or global properties aids scene categorization.

4.1 INTRODUCTION

Scene categorization is a core function of the visual system, in which the observer infers a label for the local environment. Healthy humans depend on this ability in their daily lives, using it to navigate from one place to the next and choose appropriate behaviors along the way. While scene categorization strongly depends on effectively perceiving the objects in scenes, it also utilizes information about their intrinsic global properties, such as spatial layout or size. The standard, dominant theory for scene categorization holds that these information resources are independently processed through the visual system, and only combined once they reach downstream regions responsible for decision-making (i.e. in prefrontal cortex), at which point a unified judgment of scene category is produced [74,98].

Throughout this thesis, I have described evidence for a revision to this standard scene categorization framework. In Chapters 1 and 2, we demonstrated that information about scenes' objects and spatial properties is initially combined (at least in part) as these

resources are encoded in the visual system, leading to systematic biases in both. We observed that scenes' encoded spatial properties were biased towards the values typically associated with its objects [35]. We also found that when categorizing scenes containing perceptually masked objects, participants "filled-in" those objects with features associated with the scenes' spatial properties.

Our theory is that these complementary biases aid scene categorization. This was initially speculated upon in Chapter 1, where we proposed that the object-influenced bias of scenes' spatial properties enhances their discriminability for scene classification. For instance, a "bathroom-sized" room containing an oven is perceived as more spatially similar to a kitchen than it actually is. This process supports the categorization process by driving scenes' encoded spatial properties values more typically observed to occur with their object contents.

More direct evidence for the impact of object/spatial property encoding-stage crosstalk on scene categorization was described in Chapter 2. The ability of participants to fill-in missing object information in scenes depended on first learning associations between these resources. Participants who were given the opportunity to learn these statistics were significantly more accurate at categorizing scenes with perceptually masked objects than those who did not. This means that biased object information from encoding-stage crosstalk improved scene categorization accuracy.

The studies discussed in Chapters 1 and 2 took similar approaches to infer the effect of object and spatial property crosstalk on scene categorization: scenes' object information was manipulated to control a spatial property bias or an object bias (by influencing co-occurrence statistic learning). Although these studies established

preliminary theoretical and empirical evidence that crosstalk improves scene categorization accuracy, they only addressed its impact on categorizing scenes in extreme conditions, such as when object information is obscured with perceptual masks. This leaves open the possibility that crosstalk is only evident, or perhaps brought on-line, when information about scenes' objects or spatial properties is extremely degraded. It is possible that encoding-stage crosstalk has no appreciable effect on scene categorization under typical viewing conditions.

Here we adopted a modeling approach to understand how categorization of intact (i.e. without perceptually masked content) real-world scenes is affected by crosstalk. We explored this question by creating artificial neural network (ANN) models of the visual system, and training them to categorize scenes based on information about their objects and spatial layouts. During training, one model simulated the standard framework for scene recognition and kept information about scenes' objects and spatial properties separate until it formed a category decision ("independent model"). The other model built on this basic framework with an unsupervised learning algorithm that allowed it to identify and combine co-occurring sets of object and spatial layout features before categorization ("crosstalk model").

We first validated this modeling approach by comparing scene encodings produced by each model to neural representations of those same scenes in functional magnetic resonance imaging (fMRI) recordings of human parahippocampal place area (PPA), where an encoding-stage crosstalk bias was first identified (Chapter 1). We found that both independent and crosstalk models encoded intact scenes in a way that aligned with their objective spatial properties. However, only the crosstalk model encoded scenes

as more spatially similar to the category average-size when their objects were visible versus when obscured with perceptual masks (Experiment 1). This bias emerged without being enforced and mirrored the spatial property bias previously observed in PPA and behaviorally.

These models allowed us to explore the impact of crosstalk on scene categorization under typical viewing conditions. In order to understand if crosstalk is always on-line and impacting object and spatial property encoding, we compared scene categorization decisions between human observers and both models. We found that the crosstalk model was in significantly better alignment with human decisions than the independent model was, in both typical viewing conditions, when scenes' objects were intact, and when objects in scenes were perceptually masked (Experiment 2). We also observed that the crosstalk model was more accurate than the independent model at categorizing scenes with intact objects (Experiment 3). Aligning with the findings in Chapter 2, this difference persisted when scenes' objects were obscured with perceptual masks. These results corroborate the theory that encoding-stage crosstalk of information about scenes' objects and spatial layout improves categorization, and indicate that this impact is significant even when viewing scenes as they are typically encountered in the real world.

4.2 EXPERIMENT 1

We used a modeling approach to understand how encoding-stage crosstalk between scenes' object and spatial property features affects scene recognition. One model

("independent model") resembled the standard framework for scene categorization. It processed information about the identities of objects within scenes and scenes' spatial properties into high-level features, before entering them into a classifier to produce a weighted combination of these features for scene categorization. The other model ("crosstalk model") represents a refinement of this independent model, using a similar structure while allowing for a combination of scenes' high-level object and spatial property information before the ultimate scene categorization decision is produced. Our first goal was to validate these models against neural representations of these scenes in PPA, a region of the visual system implicated in processing scenes and where evidence of crosstalk had previously been found.

We found that both models represented images of scenes with visible objects in a manner that was consistent with Humans: scene representations were ordered according to their perceived "spaciousness", a measure of its 3-Dimensional size. However, the systematic bias of scenes' spatial properties observed in PPA was only found in the crosstalk model.

4.2.1 Materials and Methods

Both models were implemented in a three-layer multilayer perceptron (MLP) algorithm, which is a basic neural network algorithm that learns a mapping from an input data source, through an intermediate hidden layer (1000 parameters) with a sigmoid activation function), to an output label. In this case, the input data source was information about scenes' objects and spatial properties, the hidden layer was configured to enable the MLP to learn a low-dimensional representation of its input (which supports categorization),

and the output was a scene category label of “bathroom”, “kitchen”, “office”, or “bedroom”. Each of these components is discussed in greater detail below.

4.2.1.1 Feature Input

We used convolutional neural nets (CNN) to extract representations of objects and spatial properties from scenes. CNNs are a class of powerful and expressive algorithms for computer vision, which learn a layered mapping from an input data source to an output label (similarly to an MLP). Importantly, representations of the input become more abstract within each successive layer of a CNN, with early layers possibly capturing edges and later layers entire objects. Indeed, these representations align with those in the visual system during passive viewing of real-world scene images [72,73,213].

While recent work has demonstrated that representations of objects in scenes naturally emerge in CNNs [214], it is unclear if spatial properties are similarly captured. We therefore trained separate CNNs to represent an operationalized version of each: one that captured local (object) features in scenes and another that captured global (spatial property) features in scenes. We enforced separate learning of these features by varying the receptive field input to each CNN. Local feature CNNs received 32x32 pixel patches from a 128x128 pixel image of a scene, whereas global feature CNNs received the entire image at once (following high-pass filtering, to promote representations aligning with spatial properties; Appendix C). This meant that in order to effectively categorize the scenes, the local CNN had to learn representations of more granular, and likely more object-centric features than the global CNN. Throughout this chapter, we refer to these local and global features as object and spatial property features, respectively.

The CNNs were implemented in MATLAB with the MatConvNet library, and were structured similarly to "AlexNet" [215]. Object and spatial property encodings of scene features were extracted from the final fully connected layer of each CNN. The values of these encodings represent normalized "activations" for each feature learned by the CNN. This means that higher-valued features have a greater likelihood of being present in an image than lower-valued ones do. More details on the CNN architecture and how they were trained can be found in Appendix C.

4.2.1.2 Model Training

The independent model and crosstalk model were trained on images of real world scenes gathered from the SUN database [216]. This dataset consisted of object and spatial property feature encodings extracted from 800 scenes in each of the following categories: bathrooms, kitchens, offices, and bedrooms (3200 total). These images were different than those used to train the CNNs (Appendix C).

Both independent and crosstalk models were trained for 1000 iterations. Over the course of a single iteration, the models received information about every scene in batches of 50. This procedure both facilitates MLP training and allowed us to implement our mechanism for object and spatial property crosstalk in the crosstalk model.

4.2.1.3 Model Structure

We developed our independent and crosstalk models of the visual system by differentiating how each received information about scenes' objects and global properties. The independent model kept these resources separate until they were pooled in a scene category classifier that represented downstream decision-making regions (i.e.

prefrontal cortex). This represents an "as-is" implementation of the MLP: it received a concatenation of scenes' object and spatial property feature encodings, pooled them in its hidden layer, and finally inferred a scene category.

In contrast, the crosstalk model allowed these resources to combine before entering the classifier. We based our instantiation of crosstalk on the experiments described in Chapter 2, where we observed that implicitly learned statistics describing co-occurring objects and spatial properties drove encoding-stage biases. During each iteration of training, the crosstalk model received information about batches of 50 scenes at a time. Every image in a batch was represented with a vector of normalized activation values that described CNN representations of its object and spatial property features. Higher values meant a higher likelihood of a given feature appearing in a scene. For instance, if the CNNs learned an object feature "shower" and a spatial property feature "small", an image of a bathroom would likely have high activations for both of these features whereas an image of a kitchen would not.

In order to estimate the co-occurrence of object and spatial property features in scenes, the model calculated Euclidean distances between the features of every batch of scenes. This yielded a distance matrix in which small distances between features meant that they likely co-occurred in that batch of images and large distances meant that they likely did not. In other words, if the batch only consisted of bathrooms, this procedure would reveal a small distance – and high similarity – between the set of object features (e.g.) corresponding to "shower" and the set of spatial property features corresponding to "small". Co-occurring feature groups were revealed with a clustering algorithm that identified ½ as many groupings as the input dimensionality. This number of groups was

not guaranteed, as only those containing both object and spatial property features were kept.

The crosstalk model performed operations on each of these co-occurring feature groups at the input layer, which gave it the opportunity to express crosstalk. It did this by combining feature values within each co-occurring object and spatial property feature group. For every scene that the crosstalk model saw, it preserved the maximum value in these co-occurring feature groups and set all others to 0, which transformed how each scene was represented. Since these features corresponded to activation values that indicated their likelihood of appearing in a given scene, high-likelihood features within every feature group dominated scene representations, while all others were shrunk to their mean. In other words, while inputs to the crosstalk model were the same dimensionality as the independent model, they exploited co-occurring features to introduce regularization, and reduce the impact of extreme feature values (i.e. noise) on their representations. This also meant that during training, the parameters of the crosstalk model learned to represent scenes in a way that emphasized these co-occurring features, giving it the opportunity express the same crosstalk biases that we observed in humans in Chapters 1 and 2.

For an example of why this mechanism works, consider an image of a bathroom, in which a large box obstructs the view of its shower. Having previously learned to associate a shower with spatial properties corresponding to the typical, average-sized bathroom, an observer's uninformative object information is reinforced with spatial property information that activates object codes consistent with the correct scene category. The crosstalk model uses a similar principle: for groups of co-occurring object

and spatial property features, it bases its decisions on scene representations that have the most informative features emphasized while regularizing all others. If the model were to learn the same association, its representation of the bathroom would have greater emphasis placed on its spatial properties than its occluded shower, yielding a representation that can be easily categorized as a bathroom. This could also lead to a systematic bias in how the model represents spatial properties. If the model viewed a different bathroom that was spatially small and had a visible shower, the association between these resources would cause the model's representation of the scene to emphasize both the scene's actual spatial properties and those associated with its shower (while regularizing unrelated features). When passed through a classifier, this representation would translate into a combination of these spatial properties, and appear biased towards the average bathroom. Both of the models are discussed in further detail in Appendix C.

4.2.1.4 Participants

Model representations were compared to neural data from the 12 human participants discussed in Chapter 1 (1 female, aged 18-23 years data [35]).

4.2.1.5 Stimuli

We used the set of bathroom images discussed in Chapter 1. These were 500 images of bathrooms gathered from the Internet that were ranked (by a set of independent raters) according to their spaciousness (Appendix A). Additional versions of the 100 most and least spacious images were produced with wavelet masks obscuring their objects that were most strongly associated with the scene category (up to three objects). In total, there

were five different types of scenes: object masked low spaciousness, object intact low spaciousness, object intact average spaciousness, object intact high spaciousness, and object masked high spaciousness. All images were also produced in a blue shade and gray shade. Human participants viewed 400 x 400 pixel images of these scenes, whereas the models received object and spatial property feature encodings from 128 x 128 pixel versions (the size was reduced to make it easier for the CNNs to extract feature encodings).

4.2.1.6 Procedure

Participants in the fMRI experiment viewed sequences of these bathroom images while having their brain activity recorded by the scanner (Chapter 1.2.2.3). Within each of 8 separate scanning runs, 252 images of the five scene types discusses above were randomly presented in blue or gray for 150ms, followed by a white fixation cross for 1350ms. For each participant, we extracted activity patterns elicited by each type of scene in right PPA, calculated their dissimilarity with Euclidean distance, and used Multidimensional Scaling (MDS) to visualize their relationships.

We used a similar procedure to visualize how the independent and crosstalk models represented these scenes, producing "synthetic participants" for each model by extracting their representations for a random sampling of bathroom images. We made 12 synthetic participants for each model, with each viewing the same number of images as human participants. We extracted encodings for every scene from each synthetic participant's hidden layer. As with the human neural data, we calculated the dissimilarity of these patterns with Euclidean distance, and visualized their relationships with MDS. To facilitate model and human comparisons, model representations were rotated with

procrustes transformations into the latent space occupied by the representations extracted from Humans.

For both Human and model representations of the five scene types, we generated bootstrapped 95% confidence intervals by resampling the averaged distance matrix across (either Human or synthetic) participants for 10,000 iterations.

4.2.2 Results

Here we validated our modeling approach to understanding how encoding-stage crosstalk between object and spatial property information impacts scene categorization. We wanted to determine if our models represented scenes similarly to humans, including the systematic bias of scenes' spatial properties observed in Chapter 1 [35].

To do this, we compared representations between the independent model, crosstalk model, and PPA for the five types of bathroom images used in the experiments discussed in Chapter 1. These images depicted bathrooms of low, average, or high spaciousness, with their objects either visible or masked. Representations of these scenes were visualized with MDS, which revealed that both of our models ordered object intact scenes according to their objective spaciousness, similar to PPA (Figure 4.1; object intact scenes are depicted with black shapes filled with a 2, 3, or 4). However, only the crosstalk model exhibited the same spatial property bias as PPA, in which extreme-sized scenes with visible objects were represented as more similar to average than when the objects in those same scenes were obscured with perceptual masks (Figure 4.1; object masked scenes are depicted with outlined shapes filled with a 1 or 5).

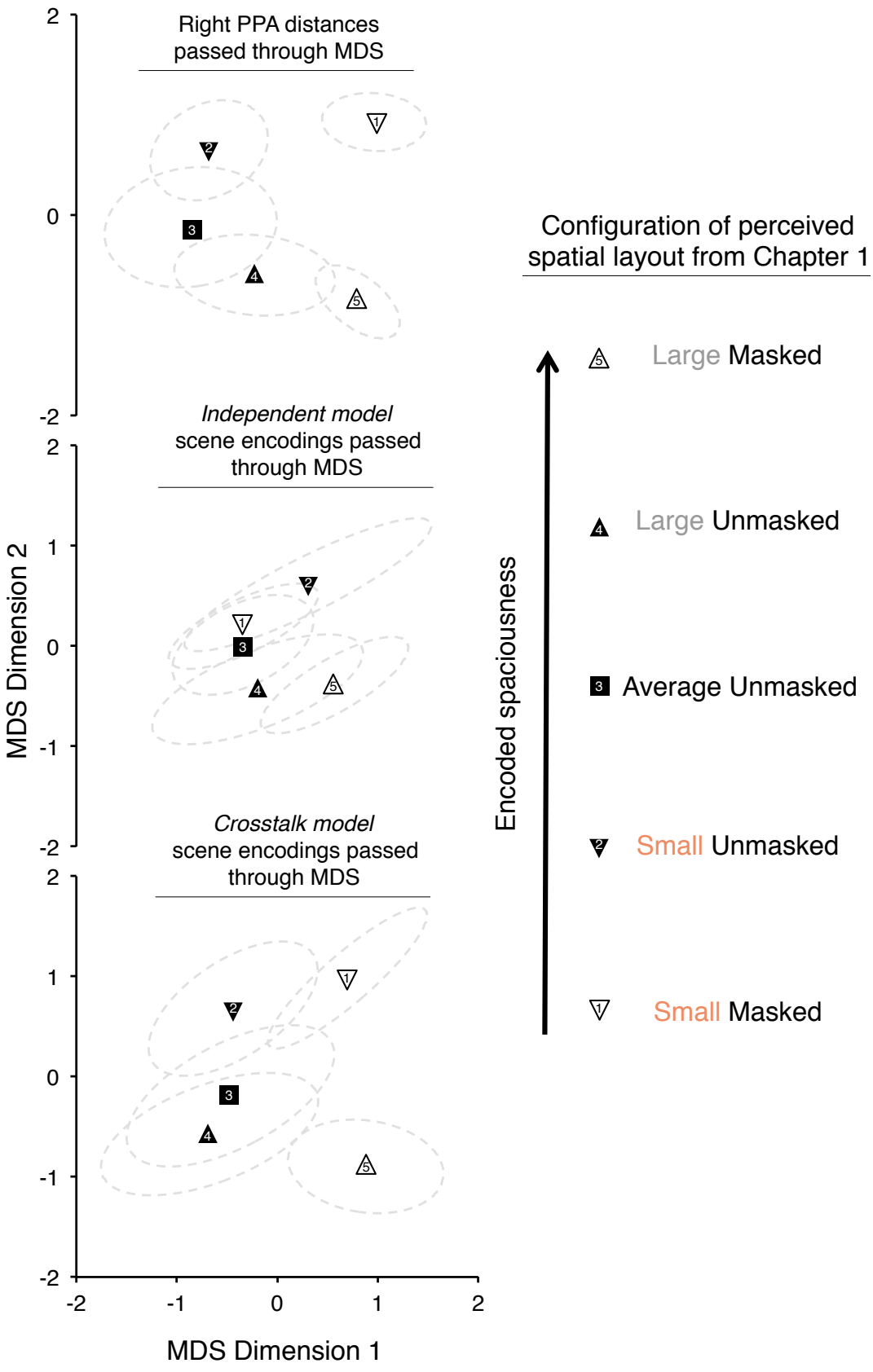
4.2.3 Experiment 1 Discussion

Comparing representations of scenes by the independent and crosstalk models and those observed in Human PPA serves two purposes. First, it validates that both models represent object intact scenes in a way that is consistent with the objective spaciousness of these scenes, as well as their representations in PPA. In other words, both models are processing information about scenes in a manner that aligns with humans.

Second, this analysis validates our crosstalk model, which we based on principles inferred from the findings of Chapter 2. The model identified and combined together groups of co-occurring object and spatial property features. It did this by emphasizing the most informative object or spatial property feature within each co-occurrence group for every scene. Since these features were normalized activation values that corresponded to their likelihood of appearing in a scene, this process simply involved preserving the maximum feature within every co-occurrence group and regularizing all others by setting them to 0. Through these principles, the model recreated the systematic bias in spatial properties observed in PPA, in which scenes were represented as more spatially similar to average when their objects were visible versus when they were perceptually masked. It must be emphasized that this bias emerged without supervision, and without any constraints specifically enforcing it. For this reason, we believe that our implementation of crosstalk is a neurally plausible one.

Figure 4.1. Spatial representations in humans and models. PPA patterns of activation

elicited by scene stimuli were visualized with MDS. Scene encodings extracted from independent and crosstalk models were similarly visualized. Although the Independent model correctly orders object visible scenes according to their relative “spaciousness”, it is not consistent with the biased configuration found in both PPA and the crosstalk model: scenes with visible objects are represented as significantly more spatially similar to average than those with masked objects. Dashed lines indicate 95% confidence intervals for each MDS plot.


Right PPA distances passed through MDS
Independent model scene encodings passed through MDS
Crosstalk model scene encodings passed through MDS
MDS Dimension 2
MDS Dimension 1
2
1
0
-1
-2
Configuration of perceived spatial layout from Chapter 1
Encoded spaciousness
Large Masked
Large Unmasked
Average Unmasked
Small Unmasked
Small Masked

4.3 EXPERIMENT 2

Having validated the independent and crosstalk models, we were next interested in using them to understand how encoding-stage crosstalk between scenes' object and spatial property information impacts scene categorization. Although our prior work has demonstrated that crosstalk can improve categorization of scenes with perceptually masked objects, the generalizability of this effect to unmasked scenes is unclear. Is it always on-line and impacting scene categorization or is it only evident in extreme cases?

We explored this question by assessing the similarity between decisions produced by human participants and our models on a four-way scene classification task of intact (object visible) scenes. We expected if crosstalk is only leveraged when scene information is extremely impoverished, then participants' decisions would correlate equally with both models. This would indicate that participants' categorization decisions are based on independently processed object and spatial property features, consistent with the standard framework for scene categorization. However, if encoding-stage crosstalk impacts scene categorization of intact scenes, we expected that humans would correlate significantly more strongly with the crosstalk model than the independent model. Our findings aligned with expectations: humans were significantly more similar to the crosstalk model than the independent model, both when objects in scenes were covered with perceptual masks, and when scenes were presented normally.

4.3.1 Materials and Methods

This experiment used the same independent and crosstalk models as discussed in Experiment 1. We compared scene categorization between Human participants and the models in three separate conditions: first, when scenes were intact (0 masked); second, when the most informative object in a scene was perceptually masked (1 masked); and third, when the two most informative objects were masked (2 masked). This allowed us to assess encoding-stage crosstalk in humans across multiple levels of information noise in scenes.

4.3.1.1 Participants

We recruited 60 participants from Amazon Mechanical Turk (20 for each of the three comparisons between models and Humans). Participants were paid $1.00 for completing the experiment, which took approximately 10 minutes.

4.3.1.2 Stimuli

This experiment used two sets of images gathered from the SUN database. The first is the set of 3200 bathroom, kitchen, office, and bedroom images described in Experiment 1. These images were used as "template images", detailed in the Procedure. We also gathered a separate set of 200 "probe images" for each of the four scene categories. This set of images did not overlap with the template images.

Human participants who completed an experiment containing intact versions of these images also provided the names of the top three objects that they most associate with each scene category. We determined the top-two most commonly named objects for

each scene category, and using the LabelMe toolbox [112], segmented their locations in every probe image. Additional versions of the probe images were created with wavelet masks covering these objects. This provided versions of every probe image with all objects intact (0 masked), one object perceptually masked (1 masked), and two objects perceptually masked (2 masked).

Both probe and template images were resized to 100 x 100 pixels for the experiments with human participants. For the models, these images were resized to 128 x 128 pixels.

4.3.1.3 Procedure

Human participants completed a web-based, four-way scene categorization task. This task was performed by using the mouse to drag-and-drop the probe image of a scene presented at the center of the screen to the template image of the same category positioned in one of the four corners of the screen. To increase task difficulty, the probe image remained visible for two seconds before it was transformed into white noise. Participants were encouraged to make their decisions as quickly as possible.

Participants were also instructed to express their confidence in their decisions through this drag-and-drop interface: the closer they placed the probe image to the template, the more confident they felt. However, we did not incorporate this information in the current analyses. Separate versions of this experiment were completed for each object masking condition (0, 1, and 2). A version of the experiment can be found at http://bit.ly/sr_model.

The independent and crosstalk models were trained to categorize the template images and then tested on the probe images. Separate tests were performed for each of the object-versions of the probe images.

4.3.1.4 Data Filtering

In light of recent evidence that time-of-completion can bias performance in behavioral experiments [169], we excluded participants who completed the experiment two standard deviations outside of average. This excluded 2 participants who completed the 0 masked object condition (18 total), 5 in the 1 masked object condition (5 total), and 4 in the 2 masked object condition (16 total).

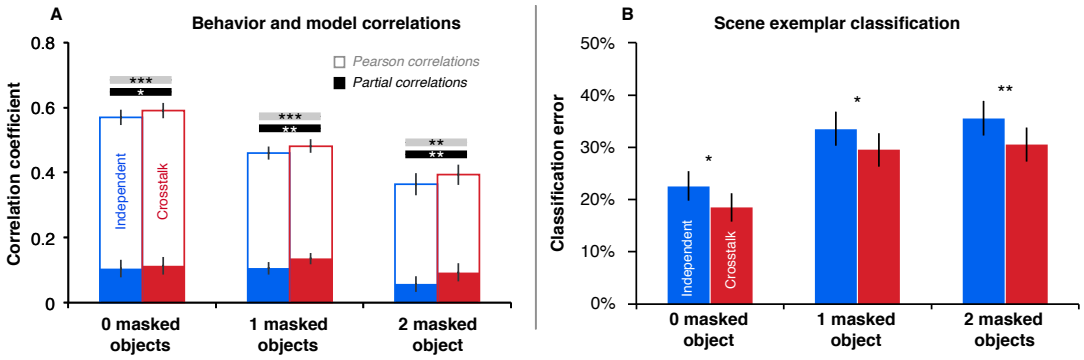


Figure 4.2. Model performance. (A) Correlations between human and model decisions on a four-way scene categorization task. Separate analyses were performed for tasks in which the image to be categorized had 0, 1, or 2 of its category-informative objects obscured with perceptual masks. Outlined bars depict Pearson correlations, while filled in bars are partial correlations (controlling for scene category label). (B) Scene categorization performance of the Independent versus Crosstalk models, calculated with percent error in classification (lower values are better). Gray bars indicate significant differences between pearson correlations or classification error; blacks bars indicate significant differences between partial correlations. All statistical comparisons were performed with 2-tailed paired samples *t*-tests. Error bars are s.e.m. * : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$.

4.3.2 Results

We compared scene categorization decisions between Human participants and both models on a four-way scene categorization task. As an initial test, we used Pearson correlations to measure the consistency between each model's decisions and every

participant's decisions when categorizing images of intact scenes (0 masked objects). We found that while both models were similar to the participants in this task, the average correlation was significantly stronger for the crosstalk model than the independent model (independent $r = 0.576$; crosstalk $r = 0.595$; $p < 0.001$).

Though this result suggests that participants leveraged encoding-stage crosstalk of scenes' object and spatial properties when categorizing intact scenes, there is an alternative explanation. It is possible this correlation simply reflects the fact that participants and the crosstalk model are both exceptionally good at categorizing scenes, and does not capture any shared underlying variability in their decision-making. To control for this possibility we used partial correlations to reassess the similarity between participants and the models after correcting for classification accuracy. Once again, we found that the crosstalk model was significantly more similar to participants than the independent model (independent $r^* = 0.104$; crosstalk $r^* = 0.113$; $p = 0.032$; Figure 4.2A 0 masked objects).

The crosstalk model was also more similar to humans than the independent model when objects in scenes were masked. Both Pearson (independent $r = 0.460$; crosstalk $r = 0.481$; $p < 0.001$) and partial correlations (independent $r^* = 0.106$; crosstalk $r^* = 0.135$; $p = 0.005$; Figure 4.2A 0 masked objects) were significantly stronger between humans and the crosstalk model when one object was masked. The pattern persisted when two objects were masked for Pearson (independent $r = 0.378$; crosstalk $r = 0.408$; $p = 0.001$) and partial correlations (independent $r^* = 0.057$; crosstalk $r^* = 0.093$; $p = 0.007$; Figure 4.2A 0 masked objects).

4.3.3 Experiment 2 Discussion

In this experiment we investigated if humans leverage encoding-stage crosstalk under typical viewing conditions. Although previous chapters in this thesis have provided theoretical and empirical evidence that crosstalk improves scene categorization, it was not clear if the effect was engaged under typical viewing conditions, such as we experience in our daily lives. Our findings indicate that it does: participants' categorization decisions for intact real world scenes were significantly more similar to the crosstalk model than the independent model, even after controlling for the classification accuracy of each.

Corroborating the findings in Chapters 1 and 2, the similarity between participants and the crosstalk model persisted as category-informative objects in the scenes were perceptually masked. Together, these results indicate that encoding-stage crosstalk is on-line and impacting scene categorization under both typical and impoverished viewing conditions.

4.4 EXPERIMENT 3

In Chapters 1 and 2, we observed systematic mutual influences of scenes' encoded object and spatial property information. We theorized that these influence worked to bring values encoded in object- and spatial property-processing pathways into alignment, thereby improving scene categorization accuracy. However, as discussed in Experiment

2, these findings were made in studies in which perceptual masks were introduced into scenes. This leaves open the possibility that object/spatial property crosstalk only works to improve scene categorization accuracy when scenes are viewed under impoverished conditions, and does not generalize to the typical viewing conditions in which we normally encounter real world scenes. Our modeling approach gave us an opportunity to answer this question by comparing scene categorization accuracy between the independent model and the crosstalk model.

4.4.1 Materials and Methods

This experiment used the same independent model and crosstalk model as Experiment 1.

4.4.1.1 Stimuli

This experiment used the same stimuli as Experiment 2.

4.4.1.2 Procedure

The independent and crosstalk models were trained (on template images) and tested (on probe images) exactly as in Experiment 2.

We tested for differences in classification accuracy between the independent model and crosstalk model using McNemar's test. This test compares the confusion matrix of each classifier and assesses the difference on a χ^2 distribution.

4.4.2 Results

We found that the crosstalk model was significantly better at categorizing object-intact scenes than the independent model (Independent error: 22.50%, Crosstalk error: 18.50%, $p = 0.027$; Figure 4.2.B). A similar result was found when one (Independent error: 33.50%, Crosstalk error: 29.50%, $p = 0.027$; Figure 4.2.B) or two (Independent error: 35.50%, Crosstalk error: 30.50%, $p = 0.009$; Figure 4.2.B) category-informative objects in the scenes were perceptually masked.

Human participants performed similarly to the crosstalk model on this task when objects in scenes were intact (Mean human error: 19.81%), when one object was masked (Mean human error: 27.60%), and when two objects were masked (Mean human error: 31.56%).

4.5 GENERAL DISCUSSION

Previous evidence for encoding-stage crosstalk between information about scenes' objects and global properties during scene categorization demonstrated that it systematically biases the encoded values of each resource. In Chapter 1, scenes' encoded spatial properties were biased towards the values associated with their category-informative objects. In Chapter 2, a mirrored version of this effect was observed: scenes' encoded objects appeared to be biased towards the values associated with their spatial properties, although this was inferred indirectly. These studies provided theoretical and

empirical evidence that these biases were not epiphenomenal, but in fact aligned and reinforced these resources in a way that improved scene categorization accuracy. However, these studies investigated encoding-stage crosstalk under viewing conditions that are not representative of the real world. The impact of object/spatial property crosstalk was measured while obscuring category-informative objects in scenes with perceptual masks. This left open the possibility that encoding-stage crosstalk only comes "on-line" when viewing extremely impoverished scenes.

In the current study we addressed this with a modeling approach. We developed an "independent" model, which was consistent with the standard framework for scene recognition, and another "crosstalk" model that allowed information about scenes' objects and spatial properties to combine before decision-making. Using these models, we found evidence that encoding-stage crosstalk impacts scene categorization under all viewing conditions: Humans categorized scenes more similarly to the crosstalk model than the independent model when their objects were intact and when they were masked. Importantly, these models also allowed us to measure how much crosstalk improves scene categorization accuracy across intact and impoverished conditions. We found that crosstalk significantly reduced categorization errors in either case.

Our crosstalk model is a simple and plausible instantiation that adds mechanisms for identifying and combining across co-occurring scene features to the standard, independent model. Intriguingly, through these simple mechanisms, it is able to recreate the spatial property bias previously found in human perceptual ratings of scene spaciousness and in patterns of activation elicited by those scenes in PPA. More research is needed to establish if this model is indeed biologically accurate, or if it is mimicking

the behavior of some other mechanism, such as recurrent computations in networks of neurons.

In summary, these experiments provide evidence that encoding-stage crosstalk between scenes' objects and spatial properties enforces biases in these resources regardless of the viewing conditions, which results in a significant and consistent boost to scene categorization accuracy.

5.0 GENERAL DISCUSSION AND CONCLUSION

The ability to identify our location in the world is essential for everyday behavior, allowing us to choose the optimal behavior for the current situation and form a plan for where to go next. Despite the ease with which we can recognize scenes, its difficulty from a technical standpoint is astonishing: a simple 32 square pixel binary image can represent approximately 4 billion unique "scenes". And yet, humans can decode the exponentially more complex retinal input into behaviorally relevant concepts like "kitchen" and "bathroom". Researchers spanning multiple fields have struggled with understanding *how* humans do this for decades.

The dominant model for scene recognition suggests that this ability emerges from independently processing complementary resources of information in scenes: their objects and their spatial properties. These resources are ultimately combined downstream of the visual system by cortical regions (such as prefrontal cortex) involved in cognition and decision-making into a scene label. In this thesis, I have presented evidence that suggests important adjustments be made to this standard model. In Chapter 1, I described evidence that scenes' perceived spatial properties are systematically biased by the identities of their objects, indicating that the visual system does not independently process these resources.

This finding motivated a revised framework for scene recognition, presented in the General Introduction. Within this framework, information about scenes' objects and spatial properties are no longer independently processed, but instead make each more consistent with their typically associated values. For instance, seeing a small rectilinear room may bias the observer to perceive it as containing a desk and computer; likewise, seeing these items in a room may bias the observer to perceive it as being spatially compact and rectilinear. Chapters 1, 2, and 3 describe how the visual system reinforces perception of each resource by leveraging knowledge of their co-occurrences. As demonstrated in Chapter 4, the enforced biases facilitate scene categorization: a simple model of the revised framework significantly improved accuracy over a model that kept these resources independent.

Theory and experimental work has described the ability of the visual system to incorporate co-occurrence statistics into perception as a canonical neural computation [163,165,217,218]. Although previous reports have described the impact of this canonical computation on low-level visual features, such as edges, the revised framework for scene recognition proposed here can be thought of as an instantiation at a higher and more abstract visual level. If it is the case that this canonical computation is applied at all levels of visual processing, it may also have a pronounced impact in other domains of information processing. For instance, it is possible that tasks such as action recognition utilize co-occurrence statistics to reinforce perception of objects and actors with the kinds of movements that they are typically associated with. More research is needed to develop and validate the general role of co-occurrence statistic learning in domains outside of scene categorization.

This thesis presents a new direction in scene recognition research, in which the effectiveness of classic, feedforward mechanisms for perception are improved with a simple and complementary form of learning. While the finding of perceptual combination represents a significant step forward in understanding scene recognition in humans, it is nonetheless incomplete along several dimensions. Future research must address its dynamics, its relationship with other forms of information processing, such as top-down feedback, and seek to understand its instantiation at a cellular level.

APPENDIX A: Encoding-stage crosstalk between object- and spatial property-based scene processing pathways

Stimulus collection and generation. Real-world scene images were 500 photographs each of bathrooms and kitchens, collected from the internet. To identify exemplars within each category that possessed spatial properties that were extreme for their category, judgments of spatial properties were obtained from 165 paid raters recruited through Amazon Mechanical Turk. Using a drag-and-drop graphical interface, each rater arranged 100 randomly selected exemplars from a single scene category along a horizontal scale according to exemplars' relative perceived "spaciousness". Each rater only worked with scenes from one of the two categories. To avoid introducing any bias into ratings, no definition of spaciousness was provided at any point. An example of the rating interface can be found at http://bit.ly/12bydKs. Raters were instructed to place the least spacious exemplars at the left end of the scale, the most spacious exemplars at the right end of the scale, and exemplars that were "about average" near the middle of the scale. Each rater judged a different set of 100 exemplars; each exemplar was judged by an average of 17 raters. Each rater's raw spaciousness judgments (i.e., left/right screen positions of scenes along the rating scale) were standardized to zero mean and unit standard deviation. Exemplars' median standardized ratings were used to arrange exemplars from each category into spaciousness quintiles.

An additional group of 61 raters judged the spaciousness of mixed sets of 50 bathrooms and 50 kitchens drawn from the middle quintiles of each category, as defined by the procedure described in the previous paragraph. Real world exemplars of both scene categories were randomly assigned to 10 equal-sized sets. Each rater's judgments of the spaciousness of all 100 rooms (50 bathrooms and 50 kitchens) within a single set were standardized as described above. Each exemplar was judged by an average of 6 raters; however, ratings from this group were only used to assess differences between the average sizes of bathrooms and kitchens, and played no role in the creation of the main experiment to follow. To allow exact specification of scenes' spatial properties and object contents, we also assembled a stimulus set composed of 500 computer-generated kitchens rendered using Trimble Sketchup (www.sketchup.com), IRender nXt 4.0 (www.renderplus.com), and custom Ruby scripts. Scenes were defined as kitchens by populating empty rooms with a refrigerator and combination stove/sink/cabinet unit. Object models were randomly selected from a pool of 10 exemplars of each object and arranged in one of six multi-object layouts. Rooms varied randomly in floor and wall covering. Variability in global spatial properties was introduced by randomly drawing the floor area of each room from a uniform distribution between 4.5 m^2 and 21.75 m^2. Rendered exemplars were divided into quintiles on the basis of floor area. As with real-world scenes images, object boundaries in scenes from high- and low-area quintiles were identified with a custom segmentation algorithm implemented in MATLAB, and an alternate set of these scenes was generated with wavelet masks obscuring the objects.

Reaction time analysis. To identify excessively long or short reactions times (RTs), data for each participant were scanned for reaction times (RTs) and trials were excluded from

analysis when RTs fell more than 4 standard deviations above or below the mean of trials accumulated across all members of his or her participant group. All operations were performed on the natural log of each RT; log transformation was conservative since it restrained strong positive skews in RT distributions, thereby making positive outliers more likely to fall within 4 standard deviations of the mean. Next, we computed the average number of trials thus excluded from each of the four adaptation runs in each experiment, and removed from further analysis all data from any run containing a number of excluded trials exceeding 4 standard deviations from the experiment average. This step was implemented to exclude runs during which participants' patterns of RT errors showed signs of relative inattentiveness. In the end, exclusions were sparing, with the percentage of trials removed by these criteria amounting to only 2.26% for the real bathroom group, 1.81% for the real kitchen group, 2.25% for the rendered group, 2.22% for the unmasked cross-category group, and 3.07% for the masked cross-category group.

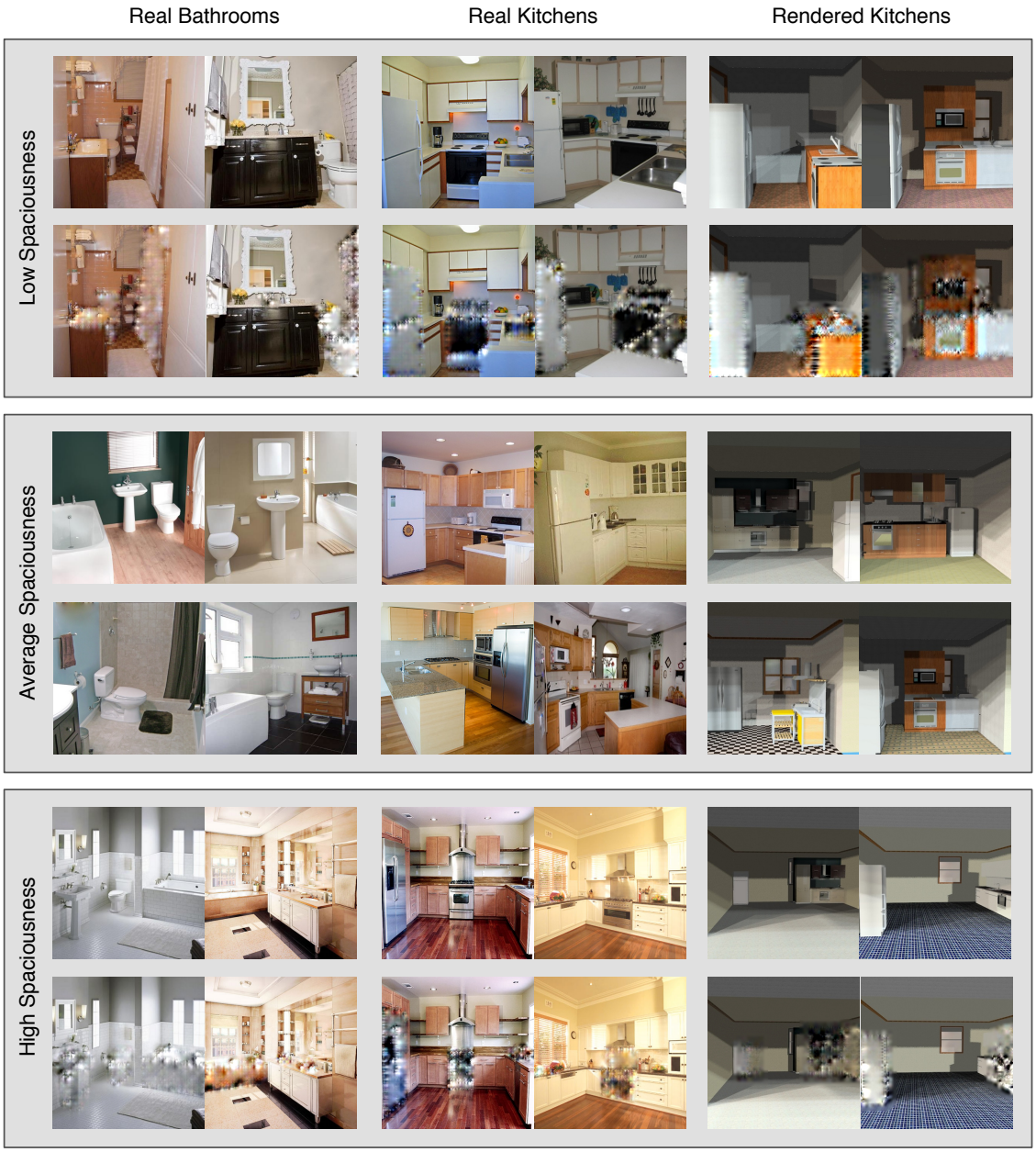
Clustering algorithm to identify inattentive participants. We noted that even though test scenes were drawn from pools encompassing the middle three quintiles of room spaciousness for each category, some participants produced long strings of identical responses that were clearly inconsistent with the spatial variability of the set of test scenes. For example, a participant in one group made identical responses to every test scene in one run, even though the average number of response transitions (i.e., number of trials in each block of 30 with a different response than the trial preceding it) across participants in that group was more than 17. Clearly, the responses of such participants were at best only weakly linked to the stimuli (likely indicative of participant inattention

to the task), and inclusion of such responses in our analyses was thus likely to cloud any adaptation effects.

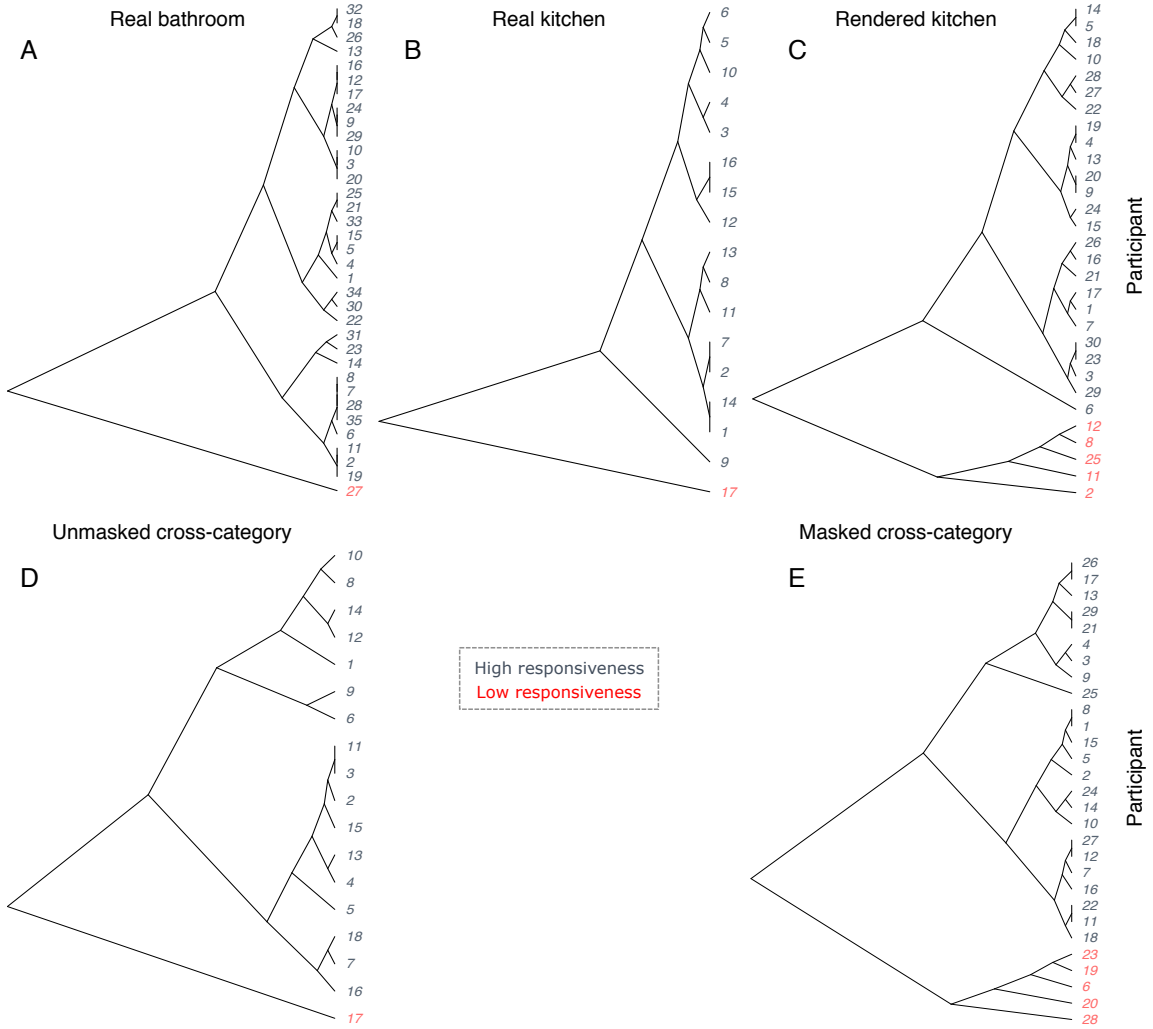
To provide an unbiased means of identifying uncooperative/inattentive participants, the responsiveness of each participant was measured by averaging the number of transitions, as defined in the preceding paragraph, across all four adaptation runs. The resulting transition scores were provided as input to an agglomerative hierarchical clustering algorithm implemented in MATLAB. Participants were divided into "high responsiveness" and "low responsiveness" clusters by cutting the resulting dendrogram for each experiment at the minimum height at which only two clusters remained. Summary statistics of judgment transitions can be found in Supplementary Table 1 and hierarchical dendrograms can be found in Supplementary Figure 2. The high responsiveness cluster in each experiment was defined as the one with the higher average number of transitions, and encompassed 34 of 35 participants in the real bathroom group, 16 of 17 participants in the real kitchen group, 25 of 30 participants in the rendered kitchen group, 17 of 18 participants in the unmasked cross-category group, and 24 of 29 of participants in the masked cross-category group. Only data from participants in high responsiveness clusters were included in further analysis.

Appendix A Table 1.
Judgment transition means and standard deviations for high- and low-responsiveness groups.

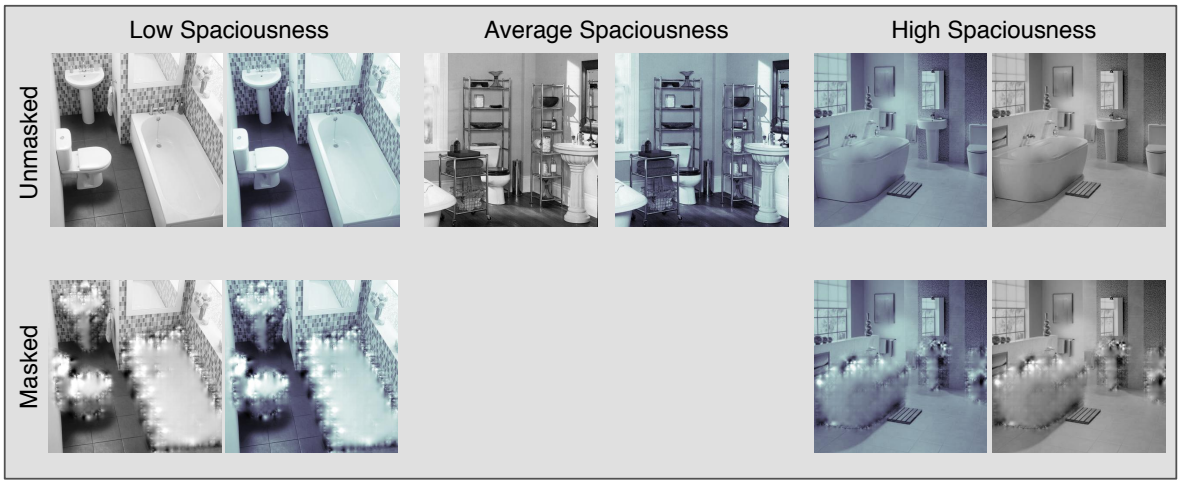
Experiment Title	High-responsiveness	Low-responsiveness
Real bathroom	20.51 (2.81)	4.00 (0.00)
Real kitchen	21.05 (1.78)	6.75 (0.00)
Rendered kitchen	19.92 (2.67)	10.80 (3.62)
Unmasked cross-category	19.32 (3.21)	6.50 (0.00)
Masked cross-category	20.22 (11.10)	11.10 (3.03)



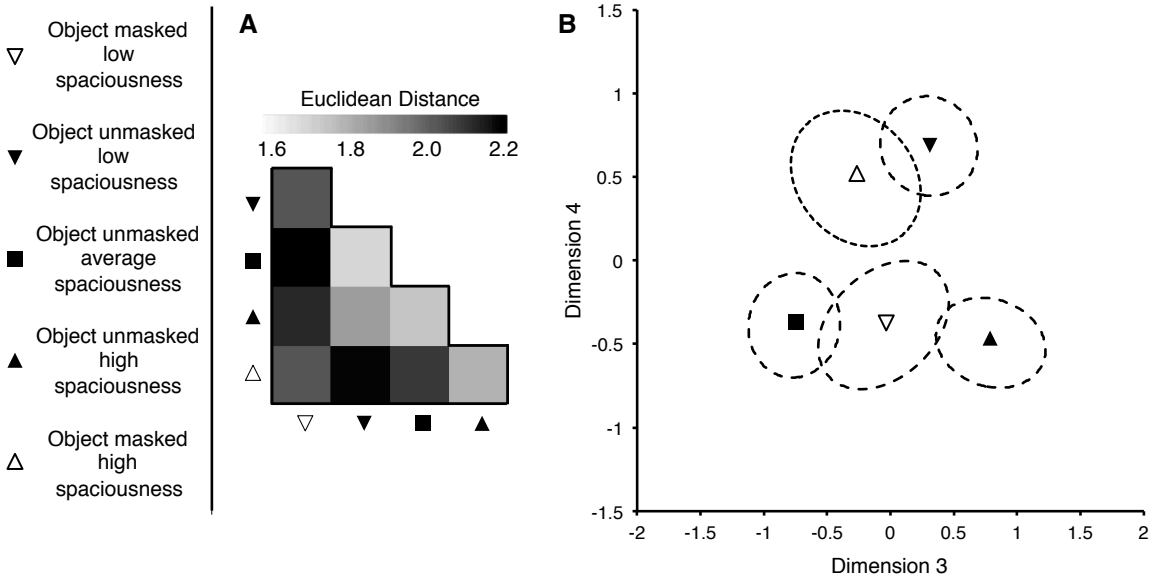
Appendix A Figure 1. Additional scene exemplars. Low and high spaciousness exemplars are shown with objects unmasked and masked.



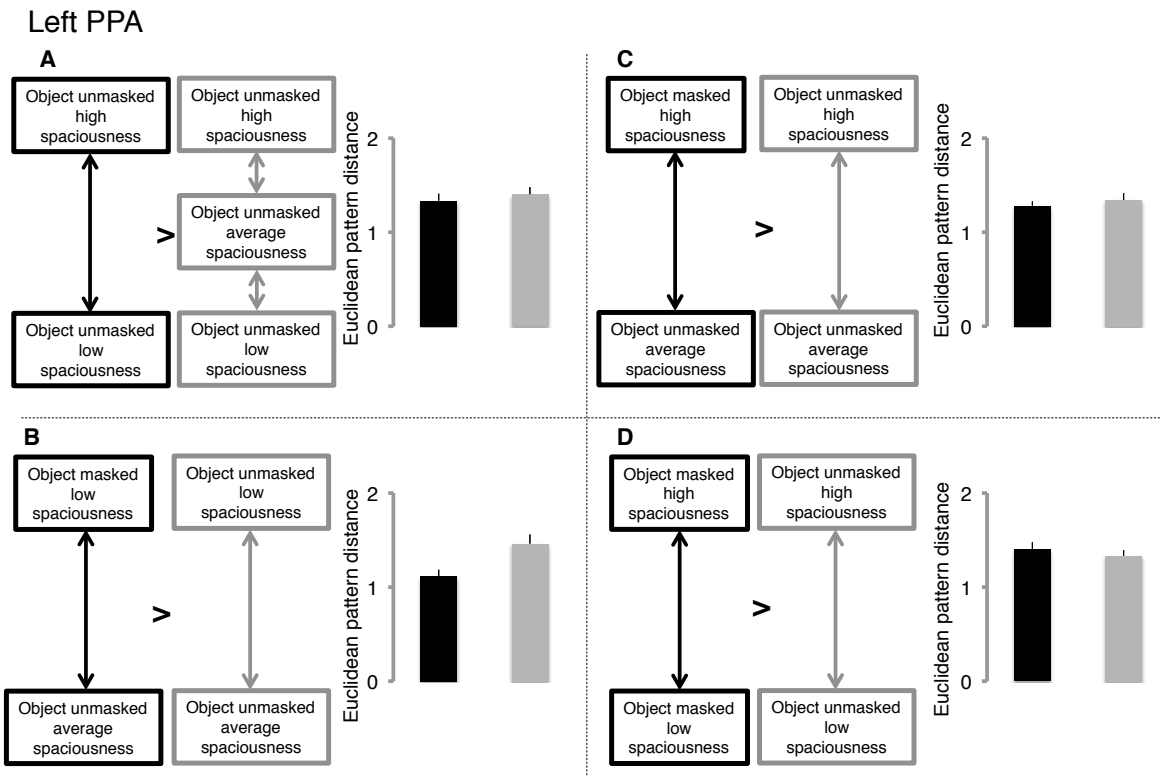
Appendix A Figure 2. High- and low-responsiveness groups extracted from behavioral experiments. For each subject, the average number of response transitions across runs was passed through agglomerative hierarchical clustering, which used a weighted average of Euclidean distance to group subjects with similar responsiveness. Each tree was cut at the shortest distance, from leaves to root, at which only two branches remained. These branches were labeled as high- and low-responsiveness groups, respectively. (a) The real bathroom high-responsiveness group consisted of 34/35 subjects; (b) Real kitchen, 16/17 subjects; (c) Rendered kitchen, 25/30 subjects; (d) Unmasked cross-category, 17/18 subjects; (e) Masked cross-category, 24/29 subjects.



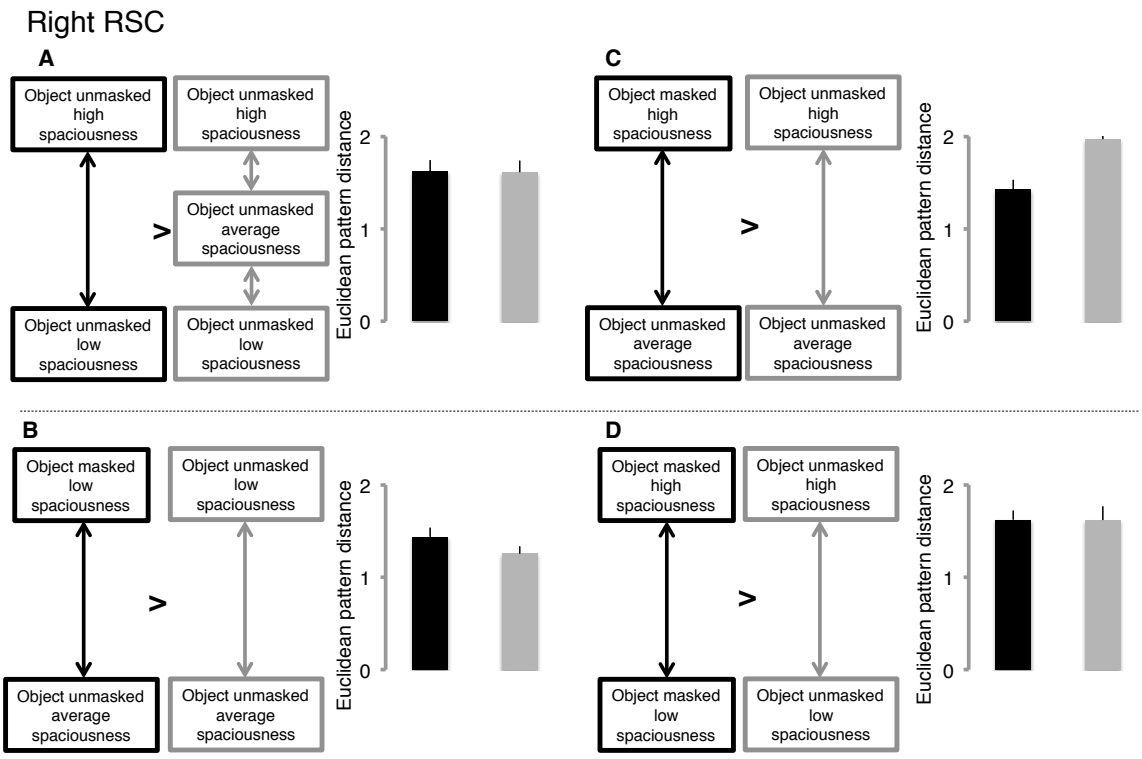
Appendix A Figure 3. Recolored scene exemplars used in the fMRI experiment.



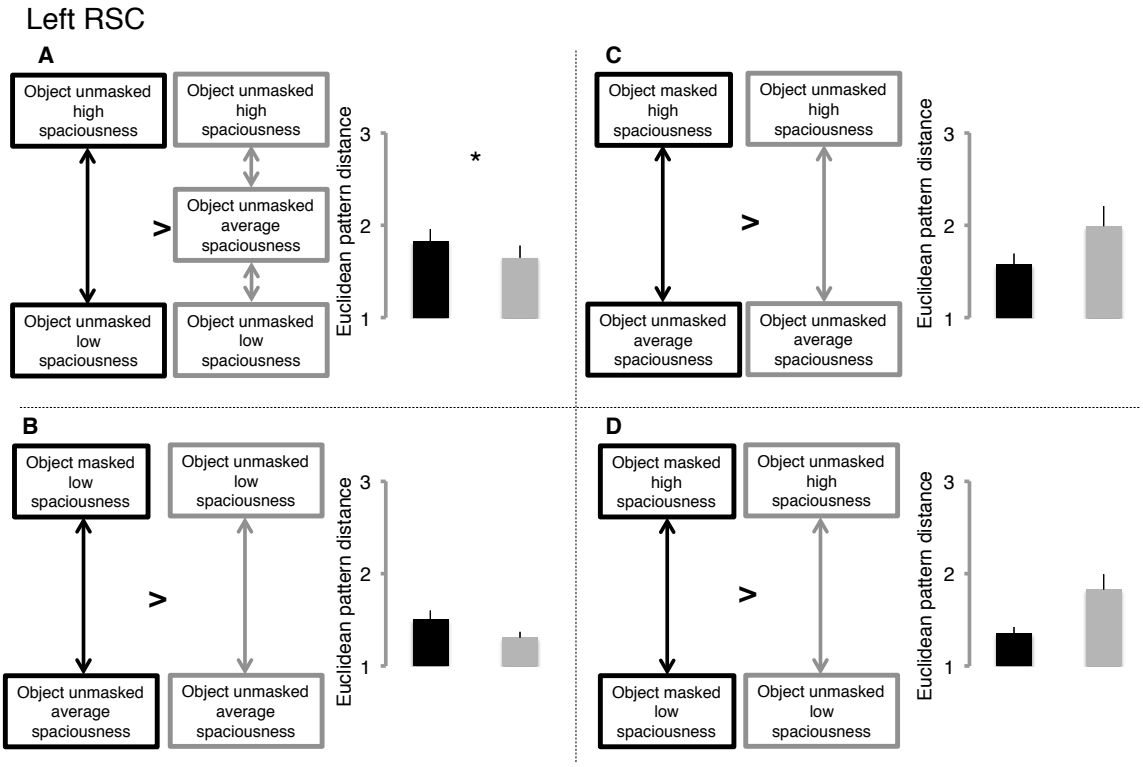
Appendix A Figure 4. Visualization of higher dimensional relationships among scene-evoked patterns in right PPA. (A) Matrix of Euclidean distances among bathroom-evoked patterns, averaged across participants (same data as Figure 4A). (B) Corresponding positions of patterns along third (horizontal) and fourth (vertical) dimensions returned by MDS. Dimensions 1 and 2 are shown in Figure 4B; dimensions 3 and 4, shown here, capture 17.5% and 15.8%, respectively, of total between-pattern distance.



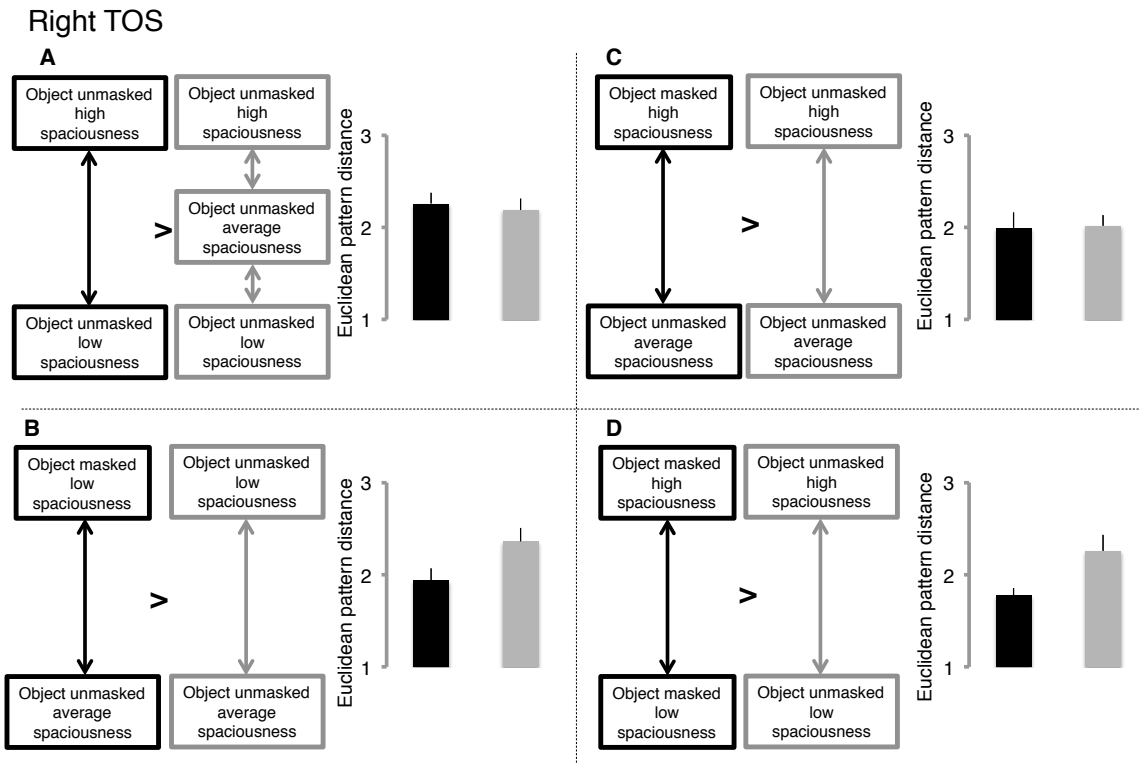
Appendix A Figure 5. Analysis of pattern distances in left PPA. (A-D) Euclidean distances between patterns of activation elicited by stimuli were not significantly different in any of the contrasts reported in the main text. Distance data in each panel correspond to comparisons between patterns denoted by same-shaded arrows in the left half of each panel. Error bars are s.e.m.



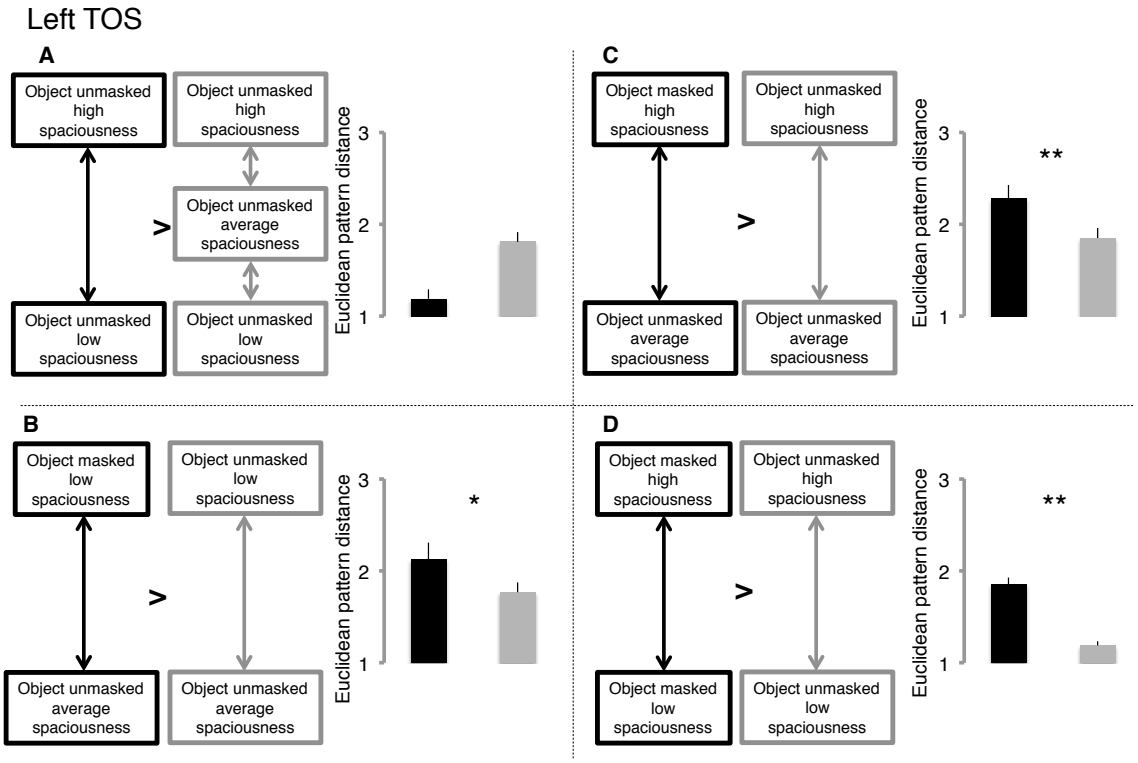
Appendix A Figure 6. Analysis of pattern distances in right RSC. (A=D) Euclidean distances between patterns of activation elicited by stimuli were not significantly different in any of the contrasts reported in the main text. Distance data in each panel correspond to comparisons between patterns denoted by same-shaded arrows in the left half of each panel. Error bars are s.e.m.



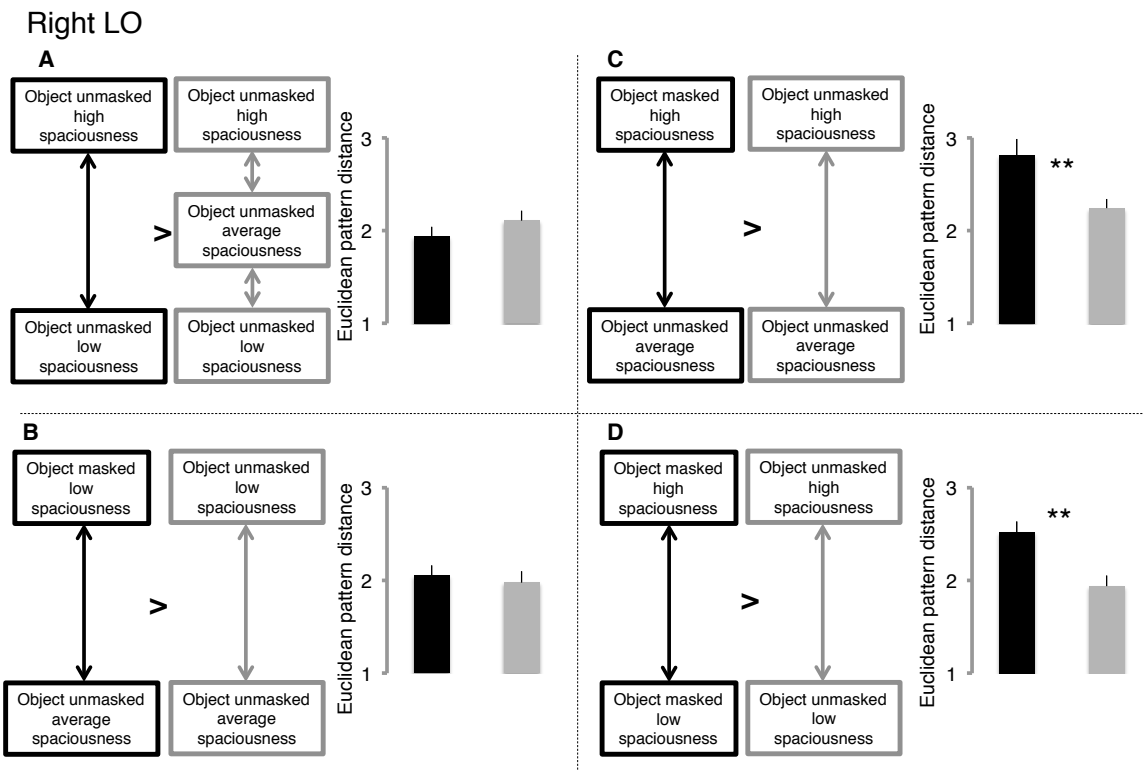
Appendix A Figure 7. Analysis of pattern distances in left RSC. (A) Average Euclidean distance between object intact high and low spaciousness exemplars was significantly different than the average of distances between each of those extremes and the pattern evoked by average-spaciousness bathrooms. (B-D) No other predicted difference in pattern distance was satisfied. Distance data in each panel correspond to comparisons between patterns denoted by same-shaded arrows in the left half of each panel. Error bars are s.e.m. *, $p < 0.05$.



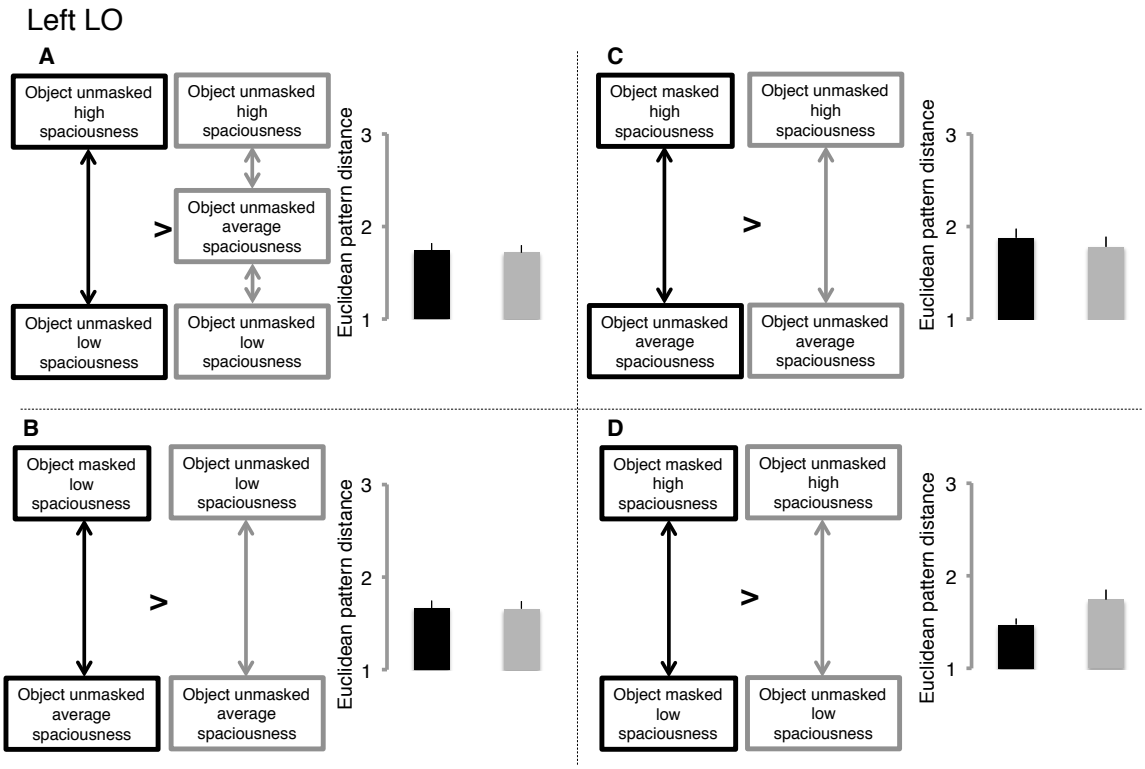
Appendix A Figure 8. Analysis of pattern distances in right TOS. (A-D) Euclidean distances between patterns of activation elicited by stimuli were not significantly different for any of the contrasts reported in the main text. Distance data in each panel correspond to comparisons between patterns denoted by same-shaded arrows in the left half of each panel. Error bars are s.e.m.



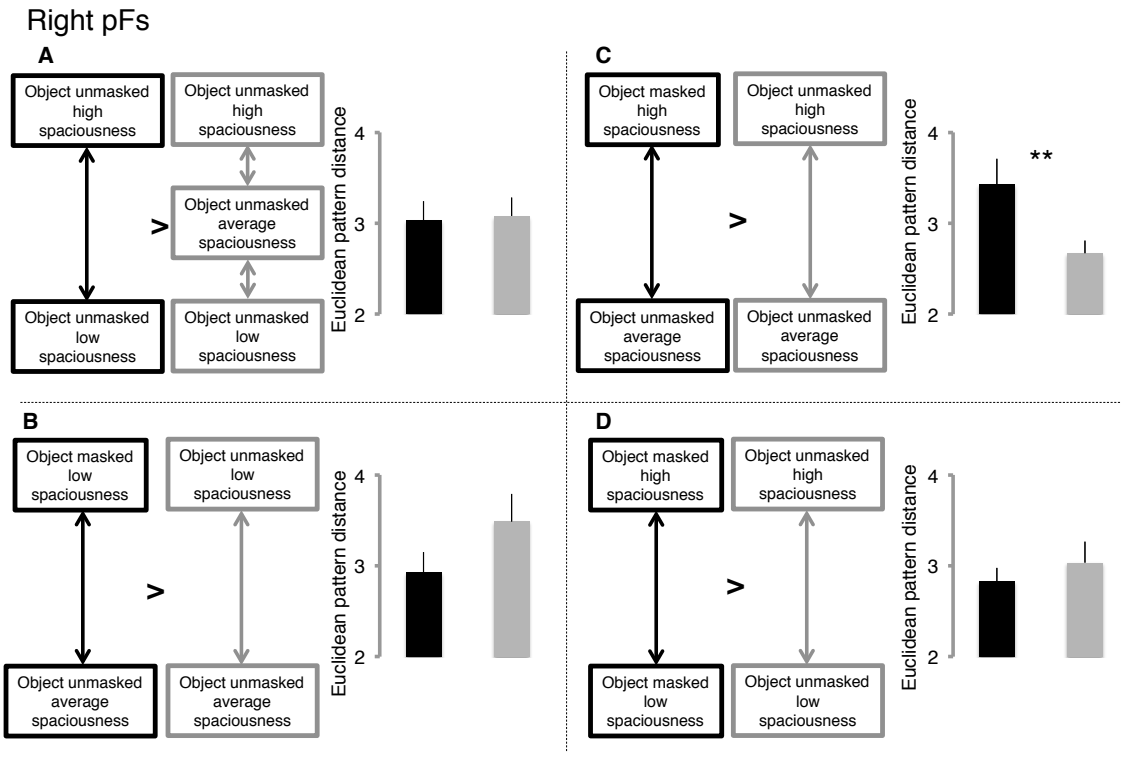
Appendix A Figure 9. Analysis of pattern distances in left TOS. (A) Average Euclidean distance between object intact high and low spaciousness exemplars was not different than the average of distances between each of those extremes and the pattern evoked by average-spaciousness bathrooms. (B-D) The remaining predicted differences in pattern distance were satisfied. Distance data in each panel correspond to comparisons between patterns denoted by same-shaded arrows in the left half of each panel. Error bars are s.e.m. *, $p < 0.05$; **, $p < 0.01$.



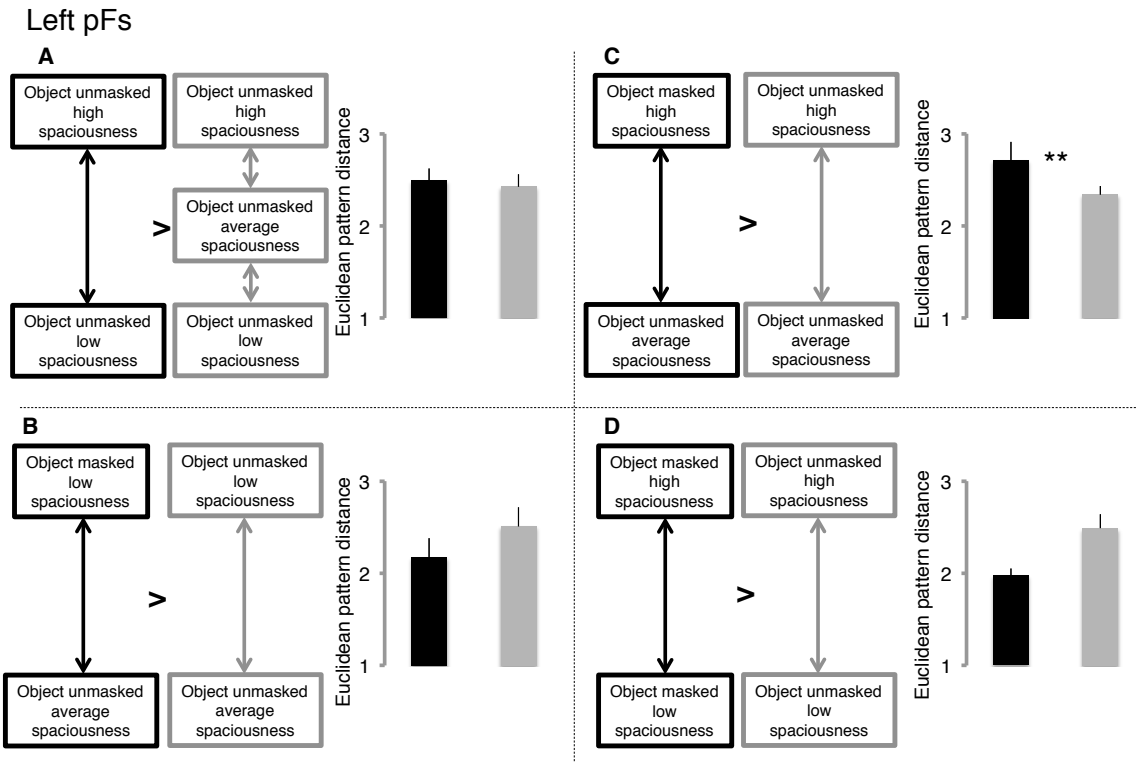
Appendix A Figure 10. Analysis of pattern distances in right LO. (A) Average Euclidean distance between object intact high and low spaciousness exemplars was not different than the average of distances between each of those extremes and the pattern evoked by average-spaciousness bathrooms. (B) Average Euclidean distance between low spaciousness and average spaciousness exemplars was not different when objects were masked versus when they were intact. (C-D) The remaining predicted differences in pattern distance were satisfied. Distance data in each panel correspond to comparisons between patterns denoted by same-shaded arrows in the left half of each panel. Error bars are s.e.m. *, $p < 0.05$; **, $p < 0.01$.



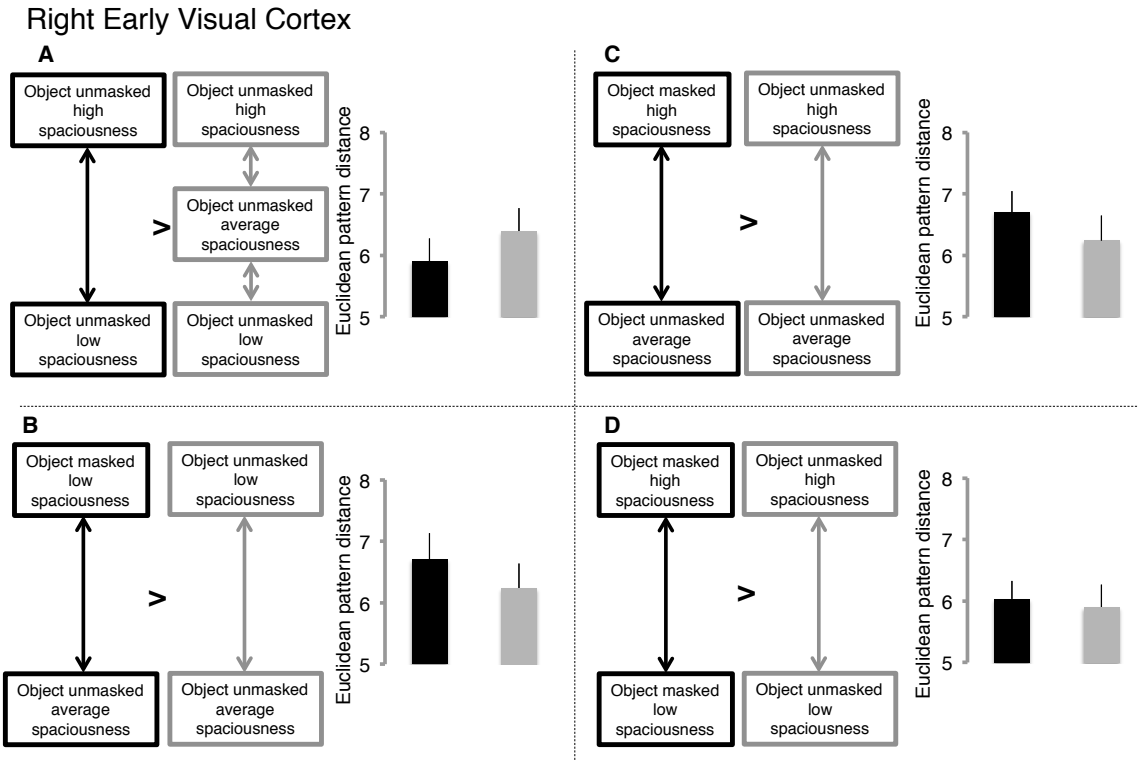
Appendix A Figure 11. Analysis of pattern distances in left LO. (A-D) Euclidean distances between patterns of activation elicited by stimuli were not significantly different in any of the contrasts reported in the main text. Distance data in each panel correspond to comparisons between patterns denoted by same-shaded arrows in the left half of each panel. Error bars are s.e.m.



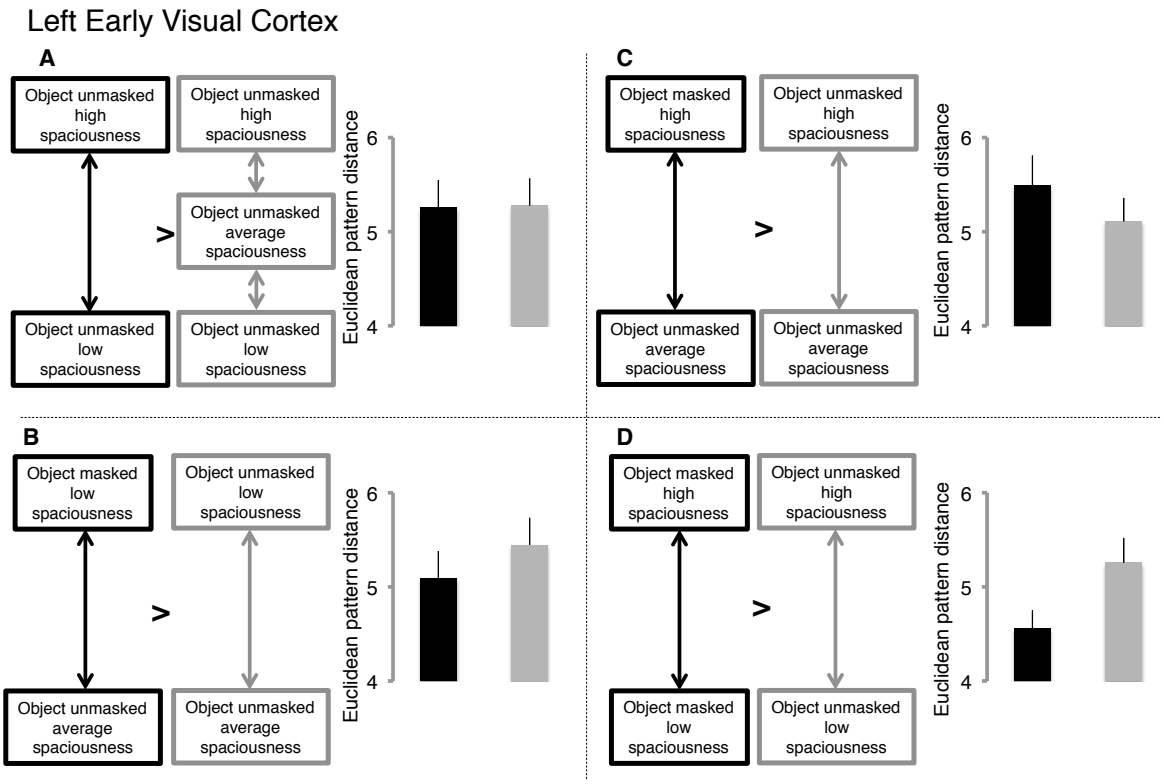
Appendix A Figure 12. Analysis of pattern distances in right pFs. (A-D) Of the contrasts, only the average Euclidean distance between high spaciousness and average spaciousness exemplars was significantly different when objects were masked versus when they were intact. No other predicted difference in pattern distance was satisfied. Distance data in each panel correspond to comparisons between patterns denoted by same-shaded arrows in the left half of each panel. Error bars are s.e.m. r**, $p < 0.01$.



Appendix A Figure 13. Analysis of pattern distances in left pFs. (A-D) Of the contrasts, only the average Euclidean distance between high spaciousness and average spaciousness exemplars was significantly different when objects were masked versus when they were intact. No other predicted difference in pattern distance was satisfied. Distance data in each panel correspond to comparisons between patterns denoted by same-shaded arrows in the left half of each panel. Error bars are s.e.m. **, $p < 0.01$.



Appendix A Figure 14. Analysis of pattern distances in right EVC. (A-D) Euclidean distances between patterns of activation elicited by stimuli were not significantly different in any of the contrasts reported in the main text. Distance data in each panel correspond to comparisons between patterns denoted by same-shaded arrows in the left half of each panel. Error bars are s.e.m.



Appendix A Figure 15. Analysis of pattern distances in right EVC. (A-D) Euclidean distances between patterns of activation elicited by stimuli were not significantly different in any of the contrasts reported in the main text. Distance data in each panel correspond to comparisons between patterns denoted by same-shaded arrows in the left half of each panel. Error bars are s.e.m.

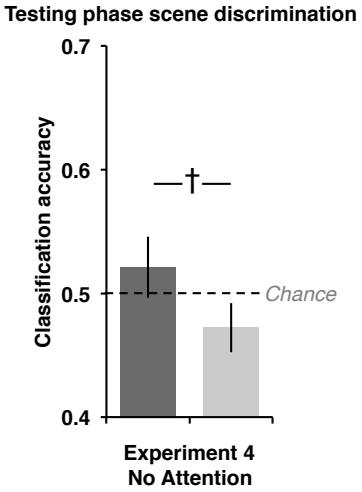
APPENDIX B: Ventral visual cortex learns object and spatial property co-occurrence statistics during scene categorization

Instructions presented during testing: “PART 1 is complete! PART 2 instructions follow:

In PART 2 you will view another slideshow of scenes. These scenes are similar to the two scene categories you saw in PART 1, where one was smaller than the other. Some of the scenes you will see are more similar to the smaller PART 1 scene category, Scene A, while others are more similar to the larger PART 1 scene category, Scene B.

After viewing each scene, you will press - (minus) if the image is like Scene A, or + (plus) if the image is like Scene B.

Please answer as quickly as possible. Press plus (+) twice or click the button in the top left twice to begin.”



Appendix B Figure 1. Experiment 4 results without controlling for attention. A separate group of 110 (61 Control) participants completed a version of Experiment 4 where they performed the color discrimination task from Experiment 1 during training. Of these participants, 89 (52 Control) passed the same data filters applied to the other experiments. Classification accuracy of testing phase exemplars for which neither group had the opportunity to learn co-occurrence statistics (i.e. those drawn from the grayscale nodes in Figure 2.2) was 52.097% for the Learning group and 47.213% for the Control group ($d = 0.361$). Error bars are s.e.m. † : $p = 0.064$.

APPENDIX C: Object and spatial property crosstalk improves scene recognition

CNN stimuli. We trained convolutional neural networks (CNNs) to discriminate between images of real-world scenes gathered from the SUN database [216]. This dataset consisted of 1151 bathrooms, 1946 kitchens, 502 offices, and 2284 bedrooms (5883 total images). We doubled the size of this dataset by including left-right flips of every image. This technique for dataset augmentation can improve the ability of a CNN to learn representations that are invariant to similar transformations.

CNN structure. CNNs were implemented in MATLAB with the MatConvNet library, and loosely followed the structure of Alexnet [215]. CNNs had 8 convolutional layers followed by a softmax classifier. The first convolutional layer filtered the 32 x 32 x 3 input image with 96 kernels of size 11 x 11 x 3 with a stride of 4 (the distance between receptive field centers of adjacent convolutional filters). The second layer applied 256 filters of size 5 x 5 x 48 (stride of 1) to the output of the first following rectification (Rectified Linear Units; ReLU), maximum response pooling (3 x 3 kernel), and normalization. This output was processed similarly to the first, before being passed through 384 filters of size 3 x 3 x 256 in layer 3. The third layer output was rectified and passed through 384 filters of size 3 x 3 x 192 in layer four. Again, these responses were rectified and passed to layer five, which contained 256 filters of the same size as layer four. This output was rectified and pooled (2 x 2 kernel), before passing through 4096 filters of size 1 x 1 x 256 in layer 6. Following rectification, 50% of the layer 6 output was set to 0 for model regularization ("dropout" to control overfitting model parameters). Layer 7 was similar, with rectification and dropout applied to the responses of its 4096 filters of size 1 x 1 x 4096. Finally, layer 8 consisted of 1000 filters of size 1 x 1 x 4096.

Local and global features of scenes were extracted from this layer, as we expected it to have the most abstract and invariant representations of these features. This layer was connected to a softmax classifier, which supported training of the network by backpropagating errors between predicted scene category labels and the true label to tune the filters in each layer.

CNN training. In order to separately capture the object and spatial property information in scene images – as was necessary to understand how crosstalk between these resources impacts categorization – we varied the receptive field input size to the models. We expected that learning about scenes with a small receptive field input would yield local features that were equivalent to object information; and learning about scenes with a large receptive field input would yield global features that were equivalent to spatial property information.

Images were initially sized to 128x128 pixels. The local feature model was trained on four (non-overlapping) 32x32 pixel patches from each image (small receptive fields). In contrast, the global feature model was trained on entire images (large receptive field).

To ensure that inputs to the global feature model were of the same size as the local feature model (so that we could use the same parameter structure for each CNN), it was trained on a wavelet-decomposed approximation of each image. This reduced the size of its input to 32x32 pixels, the same as the local feature model. We expected that applying the wavelet decomposition to images (in effect, blurring them) also encouraged

the global model to learn features that were consistent with accounts of spatial properties in humans [19,35,156,219].

Because of its receptive field size, the local feature CNN representations of scenes had four times as many dimensions than the global feature CNN. For each image, we equalized the number of local and global feature dimensions for two reasons. First, to support scene classification by reducing the dimensionality of the input to the independent and crosstalk models. Second, so that these models would not learn scene representations that were dominated local features. We did this by pooling across local features for every scene, keeping only the maximum value. This meant that every scene was represented by its most informative object features in its entire image. In the end, both local and global feature CNNs had 1000 dimensions. The correlation between local and global features, averaged across every image used to train the CNN was 0.109, indicating that each captured distinct information from scenes.

Independent and Crosstalk Models. We created two models in this experiment to investigate the impact of object and spatial property crosstalk on scene categorization: the independent model and the crosstalk model. Both models shared the same basic multilayer perceptron (MLP) structure, in which CNN features describing the objects and spatial properties in a scene image were fed into its input, then passed through a hidden layer, and finally categorized. Each of these steps, as well as the key differences between their implementation in the models is outlined below.

MLP Structure. Both models were MLPs consisting of three layers: an input layer that received concatenated object and spatial layout features, a 1000 parameter hidden layer

with a sigmoid activation function, and a softmax output layer. Models were trained by minimizing its negative log-likelihood and then backpropagating errors to optimize the values of its parameters.

MLP Input Layer. We normalized the features in every image to 0 mean and unit standard deviation before they were passed to the models. This ensured that each feature was represented by an activation, in which higher values meant a higher likelihood of its appearance in a scene image. We next normalized the values of features across all scene images to 0 mean and unit standard deviation, to standardize the distribution of activations across images for each kind of feature. This facilitated the models' ability to classify scenes. For every scene image, both models received a concatenation of its object and spatial property features from a CNN. These concatenated feature vectors were 2000 dimensions.

The key difference between the independent and crosstalk model was in their input layer. During training, both models received a batch of 50 random scene images at a time. A single iteration of training continued until the model had viewed all images in the training set. Both models had a linear transfer function from this input layer to the hidden layer.

While the independent model's input layer was an "as-is" implementation, the crosstalk model estimated co-occurring object and spatial property features within each batch. It did this by calculating the Euclidean distance between all of the feature activations across its currently viewed batch of scenes. Limiting each batch to 50 images made it easier for the model to calculate these distances and faster to train. Co-occurring

features were defined by applying a distance-based clustering algorithm (hierarchical agglomerative clustering with ward linkage) to this distance matrix. This process identified 1000 clusters, equivalent to half of the input dimensionality. However, we placed a constraint on cluster formation: each cluster had to contain at least one object and one spatial property feature. In practice this lead to the clustering algorithm identifying around 750 clusters for each batch. We made our initial choice of 1000 total clusters because it is equivalent to the original dimensionality of the object and spatial property features. Further exploration of this topic will likely yield a more optimal solution.

After identifying these clusters, the model "combined" the object and spatial property features within each. For every scene, the model iterated through these clusters and preserved the maximum activation while setting all others to 0. Since feature values were normalized activations, setting them to 0 was consistent with making the model think there was average signal for that feature. While we could have had the model set the activations to (e.g.) -3 (equivalent to 3 standard deviations below the mean), we chose to be more conservative and essentially shrink non-max features to their mean instead of introducing "negative signal".

Importantly, the ability of the crosstalk model to outperform the independent model was not solely due to the introduction of sparsity into its input. We compared the crosstalk model to an independent model that incorporated "dropout", a technique for adding sparsity at random. We set this dropout model's fraction of sparsity equivalent to the crosstalk model's, and found that it was still outperformed by the crosstalk model during the four-way scene classification test discussed in Chapter 4 (crosstalk model had

18.5% error versus the dropout model with 21% error – this pattern was similar when scenes had masked objects).

MLP Hidden Layer. Hidden layers in both models consisted of 100 parameters that captured a reduced representation of the input. Both models applied a sigmoidal function to the outputs of these neurons, which allowed it to learn non-linear representations of the input. We sampled from the hidden layer in each model for Experiment 1 in Chapter 4, in which we wanted to compare representations of scenes between human parahippocampal place area (PPA) and the models. We sampled scene information from this layer because we expected it to contain representations from each model that were high-level and could reflect the differences in each model's input layer.

MLP Output Layer. Both models had a softmax transfer function at the output layer, which converted its representations of scene images into a probability of belonging to any of the four scene categories. For every batch within each iteration, a negative log-likelihood cost was calculated between these probabilities and the true category of every scene. Errors were then backpropagated through the models to adjust its parameters to improve their ability to categorize the scenes.

BIBLIOGRAPHY

[1] Biederman I, Mezzanotte RJ, Rabinowitz JC. Scene perception: detecting and judging objects undergoing relational violations. Cognitive Psychology 1982;14:143–77.

[2] Biederman I. Recognition-by-components: a theory of human image understanding. Psychological Review 1987;94:115–47.

[3] De Graef P, Christiaens D, d'Ydewalle G. Perceptual effects of scene context on object identification. Psychol Res 1990;52:317–29.

[4] Friedman A. Framing pictures: the role of knowledge in automatized encoding and memory for gist. Journal of Experimental Psychology General 1979;108:316–55.

[5] Davenport JL, Potter MC. Scene consistency in object and background perception. Psychological Science 2004;15:559–64. doi:10.1111/j.0956-7976.2004.00719.x.

[6] Joubert OR, Rousselet GA, Fize D, Fabre-Thorpe M. Processing scene context: fast categorization and object interference. Vision Research 2007;47:3286–97. doi:10.1016/j.visres.2007.09.013.

[7] MacEvoy SP, Epstein RA. Constructing scenes from objects in human occipitotemporal cortex. Nature Neuroscience 2011;14:1323–9. doi:10.1038/nn.2903.

[8] Fei-Fei L, Iyer A, Koch C, Perona P. What do we perceive in a glance of a real-world scene? Journal of Vision 2007;7:10. doi:10.1167/7.1.10.

[9] McCotter CJ, Angle JF, Prudente LA, Mounsey JP, Ferguson JD, DiMarco JP, et al. Placement of transvenous pacemaker and ICD leads across total chronic occlusions. Pacing and Clinical Electrophysiology : PACE 2005;28:921–5. doi:10.1111/j.1540-8159.2005.00203.x.

[10] Schyns P, Oliva A. From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. Psychological Science 1994;5:195–200.

[11] Vogel J, Schwaninger A, Wallraven C, Bülthoff HH. Categorization of Natural Scenes: Local Versus Global Information and the Role of Color. ACM Trans Appl Percept 2007;4. doi:10.1145/1278387.1278393.

[12] Linsley D, MacEvoy SP. Evidence for participation by object-selective visual cortex in scene category judgments. J Vis 2014;14:19. doi:10.1167/14.9.19.

[13] Potter MC, Levy EI. Recognition memory for a rapid sequence of pictures. Journal of Experimental Psychology 1969;81:10–5. doi:10.1037/h0027470.

[14] Potter MC. Meaning in visual search. Science 1975;187:965–6.

[15] Greene MR, Oliva A. Recognition of natural scenes from global properties: Seeing the forest without representing the trees. Cogn Psychol 2009;58:137–76. doi:10.1016/j.cogpsych.2008.06.001.

[16] Greene MR, Oliva A. High-level aftereffects to global scene properties. Journal of Experimental Psychology Human Perception and Performance 2010;36:1430–42. doi:10.1037/a0019058.

[17] Greene MR, Oliva A. Natural scene categorization from conjunctions of ecological global properties. Proceedings of the 28th annual conference of the cognitive science society, 2006, p. 291–6.

[18] Renninger LW, Malik J. When is scene identification just texture recognition? Vision Research 2004;44:2301–11. doi:10.1016/j.visres.2004.04.006.

[19] Park S, Konkle T, Oliva A. Parametric Coding of the Size and Clutter of Natural Scenes in the Human Brain. Cereb Cortex 2014. doi:10.1093/cercor/bht418.

[20] Wandell B, Winawer J, Kay K. Computational modeling of responses in human visual cortex. 2014.

[21] Grill-spector K, Kourtzi Z, Kanwisher N. The lateral occipital complex and its role in object recognition. 2001.

[22] Grill-Spector K, Kushnir T, Edelman S, Avidan G, Itzchak Y, Malach R. Differential Processing of Objects under Various Viewing Conditions in the Human Lateral Occipital Complex. Neuron 1999;24:187–203. doi:10.1016/S0896-6273(00)80832-6.

[23] Malach R, Reppas JB, Benson RR, Kwong KK, Jiang H, Kennedy WA, et al. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proceedings of the National Academy of Sciences of the United States of America 1995;92:8135–9.

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

[25] Epstein R, Harris A, Stanley D, Kanwisher N. The parahippocampal place area: recognition, navigation, or encoding? Neuron 1999;23:115–25.

[26] Epstein R, Graham KS, Downing PE. Viewpoint-specific scene representations in human parahippocampal cortex. Neuron 2003;37:865–76.

[27] Avidan G, Harel M, Hendler T, Ben-Bashat D, Zohary E, Malach R. Contrast Sensitivity in Human Visual Areas and Its Relationship to Object Recognition. Journal of Neurophysiology 2002;87:3102–16.

[28] Sawamura H, Georgieva S, Vogels R, Vanduffel W, Orban GA. Using functional magnetic resonance imaging to assess adaptation and size invariance of shape processing by humans and monkeys. J Neurosci 2005;25:4294–306. doi:10.1523/JNEUROSCI.0377-05.2005.

[29] Konkle T, Oliva A. A real-world size organization of object responses in occipitotemporal cortex. Neuron 2012;74:1114–24. doi:10.1016/j.neuron.2012.04.036.

[30] Macevoy SP, Yang Z. Joint neuronal tuning for object form and position in the human lateral occipital complex. Neuroimage 2012;63:1901–8. doi:10.1016/j.neuroimage.2012.08.043.

[31] Macevoy SP. “What?” and “where?” versus “what is where?”: the impact of task on coding of object form and position in the lateral occipital complex. J Vis 2013;13. doi:10.1167/13.8.21.

[32] Kravitz DJ, Peng CS, Baker CI. Real-world scene representations in high-level visual cortex: it’s the spaces more than the places. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience 2011;31:7322–33. doi:10.1523/JNEUROSCI.4588-10.2011.

[33] Park S, Brady TF, Greene MR, Oliva A. 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 2011;31:1333–40.

[34] Cant JS, Xu Y. The Impact of Density and Ratio on Object-Ensemble Representation in Human Anterior-Medial Ventral Visual Cortex. Cereb Cortex 2014:bhu145. doi:10.1093/cercor/bhu145.

[35] Linsley D, MacEvoy SP. Encoding-Stage Crosstalk Between Object- and Spatial Property-Based Scene Processing Pathways. Cereb Cortex 2014:bhu034. doi:10.1093/cercor/bhu034.

[36] Rice GE, Watson DM, Hartley T, Andrews TJ. Low-Level Image Properties of Visual Objects Predict Patterns of Neural Response across Category-Selective Regions of the Ventral Visual Pathway. J Neurosci 2014;34:8837–44. doi:10.1523/JNEUROSCI.5265-13.2014.

[37] Walther DB, Chai B, Caddigan E, Beck DM, Fei-Fei L. Simple line drawings suffice for functional MRI decoding of natural scene categories. PNAS 2011;108:9661–6. doi:10.1073/pnas.1015666108.

[38] Park S, Chun MM. Different roles of the parahippocampal place area (PPA) and retrosplenial cortex (RSC) in panoramic scene perception. Neuroimage 2009;47:1747–56. doi:10.1016/j.neuroimage.2009.04.058.

[39] Harel A, Kravitz DJ, Baker CI. Deconstructing Visual Scenes in Cortex: Gradients of Object and Spatial Layout Information. Cereb Cortex 2013;23:947–57. doi:10.1093/cercor/bhs091.

[40] Epstein RA. Parahippocampal and retrosplenial contributions to human spatial navigation. Trends Cogn Sci (Regul Ed) 2008;12:388–96. doi:10.1016/j.tics.2008.07.004.

[41] Epstein RA, Vass LK. Neural systems for landmark-based wayfinding in humans. Phil Trans R Soc B 2014;369:20120533. doi:10.1098/rstb.2012.0533.

[42] Marchette SA, Vass LK, Ryan J, Epstein RA. Anchoring the neural compass: coding of local spatial reference frames in human medial parietal lobe. Nat Neurosci 2014;advance online publication. doi:10.1038/nn.3834.

[43] Nakamura K, Kawashima R, Sato N, Nakamura A, Sugiura M, Kato T, et al. Functional delineation of the human occipito-temporal areas related to face and scene processing. A PET study. Brain 2000;123 (Pt 9):1903–12.

[44] Grill-Spector K. The neural basis of object perception. Curr Opin Neurobiol 2003;13:159–66.

[45] Hasson U, Harel M, Levy I, Malach R. Large-scale mirror-symmetry organization of human occipito-temporal object areas. Neuron 2003;37:1027–41.

[46] MacEvoy SP, Epstein RA. Position selectivity in scene- and object-responsive occipitotemporal regions. J Neurophysiol 2007;98:2089–98. doi:10.1152/jn.00438.2007.

[47] Dilks DD, Julian JB, Kubilius J, Spelke ES, Kanwisher N. Mirror-image sensitivity and invariance in object and scene processing pathways. J Neurosci 2011;31:11305–12. doi:10.1523/JNEUROSCI.1935-11.2011.

[48] Konkle T, Caramazza A. Tripartite Organization of the Ventral Stream by Animacy and Object Size. J Neurosci 2013;33:10235–42. doi:10.1523/JNEUROSCI.0983-13.2013.

[49] Epstein RA, Morgan LK. Neural responses to visual scenes reveals inconsistencies between fMRI adaptation and multivoxel pattern analysis. Neuropsychologia 2012;50:530–43. doi:10.1016/j.neuropsychologia.2011.09.042.

[50] Bettencourt KC, Xu Y. The role of transverse occipital sulcus in scene perception and its relationship to object individuation in inferior intraparietal sulcus. J Cogn Neurosci 2013;25:1711–22. doi:10.1162/jocn_a_00422.

[51] Baldassano C, Beck DM, Fei-Fei L. Differential connectivity within the Parahippocampal Place Area. Neuroimage 2013;75:228–37. doi:10.1016/j.neuroimage.2013.02.073.

[52] Aguirre GK, Zarahn E, D'Esposito M. An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. Neuron 1998;21:373–83.

[53] Aguirre GK, Detre JA, Alsop DC, D'Esposito M. The Parahippocampus Subserves Topographical Learning in Man. Cereb Cortex 1996;6:823–9. doi:10.1093/cercor/6.6.823.

[54] Milner AD, Perrett DI, Johnston RS, Benson PJ, Jordan TR, Heeley DW, et al. Perception and action in "visual form agnosia." Brain 1991;114 (Pt 1B):405–28.

[55] James TW, Culham J, Humphrey GK, Milner AD, Goodale MA. Ventral occipital lesions impair object recognition but not object-directed grasping: an fMRI study. Brain 2003;126:2463–75. doi:10.1093/brain/awg248.

[56] Karnath H-O, Rüter J, Mandler A, Himmelbach M. The Anatomy of Object Recognition—Visual Form Agnosia Caused by Medial Occipitotemporal Stroke. J Neurosci 2009;29:5854–62. doi:10.1523/JNEUROSCI.5192-08.2009.

[57] McIntosh RD, Dijkerman HC, Mon-Williams M, Milner AD. Grasping What is Graspable: Evidence from Visual form Agnosia. Cortex 2004;40:695–702. doi:10.1016/S0010-9452(08)70165-5.

[58] Dijkerman HC, Milner AD, Carey DP. Grasping Spatial Relationships: Failure to Demonstrate Allocentric Visual Coding in a Patient with Visual Form Agnosia. Consciousness and Cognition 1998;7:424–37. doi:10.1006/ccog.1998.0365.

[59] Walsh V, Cowey A. Transcranial magnetic stimulation and cognitive neuroscience. Nat Rev Neurosci 2000;1:73–80. doi:10.1038/35036239.

[60] Mueller JK, Grigsby EM, Prevosto V, Petraglia Iii FW, Rao H, Deng Z-D, et al. Simultaneous transcranial magnetic stimulation and single-neuron recording in alert non-human primates. Nat Neurosci 2014;17:1130–6. doi:10.1038/nn.3751.

[61] Dilks DD, Julian JB, Paunov AM, Kanwisher N. The occipital place area (OPA) is causally and selectively involved in scene perception. J Neurosci 2013;33:1331–136a. doi:10.1523/JNEUROSCI.4081-12.2013.

[62] Julian JB, Ryan J, Hamilton RH, Epstein RA. The Occipital Place Area Is Causally Involved in Representing Environmental Boundaries during Navigation. Current Biology 2016;0. doi:10.1016/j.cub.2016.02.066.

[63] Mullin CR, Steeves JKE. TMS to the lateral occipital cortex disrupts object processing but facilitates scene processing. J Cogn Neurosci 2011;23:4174–84. doi:10.1162/jocn_a_00095.

[64] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 1943;5:115–33. doi:10.1007/BF02478259.

[65] Hebb DO. The organization of behavior: a neuropsychological theory. Wiley; 1949.

[66] Hinton G. The ups and downs of Hebb synapses. Canadian Psychology/Psychologie Canadienne 2003;44:10–3. doi:10.1037/h0085812.

[67] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Object Detectors Emerge in Deep Scene CNNs. arXiv:14126856 [cs] 2014.

[68] Olshausen BA, Field DJ. Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Research 1997;37:3311–25. doi:10.1016/S0042-6989(97)00169-7.

[69] Stansbury DE, Naselaris T, Gallant JL. Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex. Neuron 2013;79:1025–34. doi:10.1016/j.neuron.2013.06.034.

[70] Agrawal P, Stansbury D, Malik J, Gallant JL. Pixels to Voxels: Modeling Visual Representation in the Human Brain. arXiv:14075104 [cs, Q-Bio] 2014.

[71] Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, et al. Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. arXiv:14063284 [cs, Q-Bio] 2014.

[72] Guclu U, van Gerven MAJ. Unsupervised Feature Learning Improves Prediction of Human Brain Activity in Response to Natural Images. PLoS Comput Biol 2014;10. doi:10.1371/journal.pcbi.1003724.

[73] Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. PNAS 2014;111:8619–24. doi:10.1073/pnas.1403112111.

[74] Hollingworth A, Henderson JM. Object identification is isolated from scene semantic constraint: evidence from object type and token discrimination. Acta Psychol (Amst) 1999;102:319–43.

[75] Freedman DJ, Riesenhuber M, Poggio T, Miller EK. Visual Categorization and the Primate Prefrontal Cortex: Neurophysiology and Behavior. Journal of Neurophysiology 2002;88:929–41.

[76] Biederman I. Recognition-by-components: a theory of human image understanding. Psychological Review 1987;94:115–47.

[77] Biederman I, Mezzanotte RJ, Rabinowitz JC. Scene perception: detecting and judging objects undergoing relational violations. Cognitive Psychology 1982;14:143–77.

[78] De Graef P, Christiaens D, d'Ydewalle G. Perceptual effects of scene context on object identification. Psychological Research 1990;52:317–29. doi:10.1007/BF00868064.

[79] Friedman A. Framing Pictures: The Role of Knowledge in Automatized Encoding and Memory for Gist. Journal of Experimental Psychology 1979;108:316–55.

[80] Biederman I. Perceiving real-world scenes. Science 1972;177:77–80.

[81] MacEvoy SP, Epstein RA. Constructing scenes from objects in human occipitotemporal cortex. Nature Neuroscience 2011;14:1323–9.

[82] Davenport JL, Potter MC. Scene Consistency in Object and Background Perception. Psychological Science 2004;15:559–64.

[83] Joubert OR, Rousselet GA, Fize D, Fabre-Thorpe M. Processing scene context: Fast categorization and object interference. Vision Research 2007;47:3286–97. doi:10.1016/j.visres.2007.09.013.

[84] Fei-Fei L, Iyer A, Koch C, Perona P. What do we perceive in a glance of a real-world scene? Journal of Vision 2007;7:10,1–29.

[85] Greene MR, Oliva A. Recognition of natural scenes from global properties: seeing the forest without representing the trees. Cognitive Psychology 2009;58:137–76.

[86] McCotter M, Gosselin F, Sowden P, Schyns P. The use of visual information in natural scenes. Visual Cognition 2005;12:938–53. doi:10.1080/13506280444000599.

[87] Oliva A, Schyns PG. Diagnostic colors mediate scene recognition. Cognitive Psychology 2000;41:176–210.

[88] Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 2001;42:145–75.

[89] Oliva A, Torralba A. Building the gist of a scene: the role of global image features in recognition. Progress in Brain Research 2006;155:23–36.

[90] Renninger LW, Malik J. When is scene identification just texture recognition? Vision Research 2004;44:2301–11.

[91] Schyns PG, Oliva A. From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. Psychological Science 1994;5:195–200.

[92] Vogel J, Schiele B. Semantic Modeling of Natural Scenes for Content-Based Image Retrieval. International Journal of Computer Vision 2007;72:133–57. doi:10.1007/s11263-006-8614-1.

[93] Palmer SE. The effects of contextual scenes on the identification of objects. Memory & Cognition 1975;3:519–26. doi:10.3758/BF03197524.

[94] Antes JR, Penland JG, Metzger RL. Processing global information in briefly presented pictures. Psychological Research 1981;43:277–92. doi:10.1007/BF00308452.

[95] Loftus GR, Nelson WW, Kallman HJ. Differential acquisition rates for different types of information from pictures. The Quarterly Journal of Experimental Psychology Section A 1983;35:187–98. doi:10.1080/14640748308402124.

[96] Boyce SJ, Pollatsek A. Identification of objects in scenes: the role of scene background in object naming. J Exp Psychol Learn Mem Cogn 1992;18:531–43.

[97] Bar M, Ullman S. Spatial context in recognition. Perception 1996;25:343–52.

[98] Hollingworth A, Henderson JM. Does consistent scene context facilitate object perception? J Exp Psychol Gen 1998;127:398–415.

[99] Bar M. Visual objects in context. Nature Reviews Neuroscience 2004;5:617–29.

[100] Mudrik L, Lamy D, Deouell LY. ERP evidence for context congruity effects during simultaneous object–scene processing. Neuropsychologia 2010;48:507–17. doi:10.1016/j.neuropsychologia.2009.10.011.

[101] Greene MR, Oliva A. High-level aftereffects to global scene properties. Journal of Experimental Psychology: Human Perception and Performance 2010;36:1430–42. doi:10.1037/a0019058.

[102] Aguirre GK, Zarahn E, D'Esposito M. An area within human ventral cortex sensitive to "building" stimuli: Evidence and implications. Neuron 1998;21:373–83.

[103] Epstein RA, Kanwisher N. A cortical representation of the local visual environment. Nature 1998;392:598–601.

[104] Kravitz DJ, Peng CS, Baker CI. Real-World Scene Representations in High-Level Visual Cortex: It's the Spaces More Than the Places. The Journal of Neuroscience 2011;31:7322–33. doi:10.1523/JNEUROSCI.4588-10.2011.

[105] Mullally SL, Maguire EA. A New Role for the Parahippocampal Cortex in Representing Space. The Journal of Neuroscience 2011;31:7441–9. doi:10.1523/JNEUROSCI.0267-11.2011.

[106] Mullally SL, Maguire EA. Exploring the role of space-defining objects in constructing and maintaining imagined scenes. Brain Cogn 2013;82:100–7. doi:10.1016/j.bandc.2013.02.013.

[107] Park S, Brady TF, Greene MR, Oliva A. 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 2011;31:1333–40. doi:10.1523/JNEUROSCI.3885-10.2011.

[108] Aminoff E, Gronau N, Bar M. The Parahippocampal Cortex Mediates Spatial and Nonspatial Associations. Cerebral Cortex 2007;17:1493–503.

[109] Hassabis D, Kumaran D, Maguire EA. Using Imagination to Understand the Neural Basis of Episodic Memory. J Neurosci 2007;27:14365–74. doi:10.1523/JNEUROSCI.4549-07.2007.

[110] Summerfield JJ, Hassabis D, Maguire EA. Differential engagement of brain regions within a “core” network during scene construction. Neuropsychologia 2010;48:1501–9. doi:10.1016/j.neuropsychologia.2010.01.022.

[111] Howard LR, Kumaran D, Ólafsdóttir HF, Spiers HJ. Double Dissociation between Hippocampal and Parahippocampal Responses to Object–Background Context and Scene Novelty. J Neurosci 2011;31:5253–61. doi:10.1523/JNEUROSCI.6055-10.2011.

[112] Russell B, Torralba A, Murphy K, Freeman W. LabelMe: A Database and Web-Based Tool for Image Annotation. International Journal of Computer Vision 2008;77:157–73. doi:10.1007/s11263-007-0090-8.

[113] Honey C, Kirchner H, VanRullen R. Faces in the cloud: Fourier power spectrum biases ultrarapid face detection. J Vis 2008;8. doi:10.1167/8.12.9.

[114] Brainard DH. The Psychophysics Toolbox. Spatial Vision 1997;10:433–6.

[115] Rhodes G, Jeffery L, Clifford CWG, Leopold DA. The timecourse of higher-level face aftereffects. Vision Research 2007;47:2291–6. doi:10.1016/j.visres.2007.05.012.

[116] Greene MR, Oliva A. High-level aftereffects to global scene properties. J Exp Psychol Hum Percept Perform 2010;36:1430–42. doi:10.1037/a0019058.

[117] Aguirre GK, Mattar MG, Magis-Weinberg L. de Bruijn cycles for neural decoding. NeuroImage 2011;56:1293–300. doi:10.1016/j.neuroimage.2011.02.005.

[118] MacEvoy SP, Yang Z. Joint neuronal tuning for object form and position in the human lateral occipital complex. NeuroImage 2012;63:1901–8.

[119] Epstein R, Higgins J. Differential parahippocampal and retrosplenial involvement in three types of visual scene recognition. Cerebral Cortex 2006;17:1680–93.

[120] Drucker DM, Aguirre GK. Different spatial scales of shape similarity representation in lateral and ventral LOC. Cereb Cortex 2009;19:2269–80.

[121] Morgan LK, MacEvoy SP, Aguirre GK, Epstein RA. Distances between real-world locations are represented in the human hippocampus. The Journal of Neuroscience 2011;31:1238–45. doi:Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S.

[122] Kriegeskorte N, Goebel R, Bandettini P. Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America 2006;103:3863–8. doi:10.1073/pnas.0600244103.

[123] Nichols TE, Holmes AP. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Human Brain Mapping 2002;15:1–25.

[124] Julian JB, Fedorenko E, Webster J, Kanwisher N. An algorithmic method for functionally defining regions of interest in the ventral visual pathway. NeuroImage 2012;60:2357–64. doi:10.1016/j.neuroimage.2012.02.055.

[125] Webster M, Maclin O. Figural aftereffects in the perception of faces. Psychonomic Bulletin & Review 1999;6:647–53. doi:10.3758/BF03212974.

[126] Leopold DA, O'Toole AJ, Vetter T, Blanz V. Prototype-referenced shape encoding revealed by high-level aftereffects. Nature Neuroscience 2001;4:89–94. doi:10.1038/82947.

[127] Webster MA, Kaping D, Mizokami Y, Duhamel P. Adaptation to natural facial categories. Nature 2004;428:557–61. doi:10.1038/nature02420.

[128] Little AC, DeBruine LM, Jones BC. Sex-contingent face after-effects suggest distinct neural populations code male and female faces. Proc R Soc B 2005;272:2283–7. doi:10.1098/rspb.2005.3220.

[129] Harel A, Kravitz DJ, Baker CI. Deconstructing Visual Scenes in Cortex: Gradients of Object and Spatial Layout Information. Cerebral Cortex 2012:DOI:10.1093/cercor/bhs091. doi:10.1093/cercor/bhs091.

[130] Epstein RA, Graham KS, Downing PE. Viewpoint-specific scene representations in human parahippocampal cortex. Neuron 2003;37:865–76.

[131] Kirchhoff BA, Wagner AD, Maril A, Stern CE. Prefrontal-temporal circuitry for episodic encoding and subsequent memory. Journal of Neuroscience 2000;20:6173–80.

[132] Wagner AD, Schacter DL, Rotte M, Koutstaal W, Maril A, Dale AM, et al. Building memories: remembering and forgetting of verbal experiences as predicted by brain activity. [see comments.]. Science 1998;281:1188–91.

[133] Rajimehr R, Devaney KJ, Bilenko NY, Young JC, Tootell RBH. The "Parahippocampal Place Area" Responds Preferentially to High Spatial Frequencies in Humans and Monkeys. PLoS Biol 2011;9. doi:10.1371/journal.pbio.1000608.

[134] Zimmer M, Kovács G. Position specificity of adaptation-related face aftereffects. Phil Trans R Soc B 2011;366:586–95. doi:10.1098/rstb.2010.0265.

[135] Anstis S, Verstraten FA., Mather G. The motion aftereffect. Trends in Cognitive Sciences 1998;2:111–7. doi:10.1016/S1364-6613(98)01142-5.

[136] Zeidman P, Mullally SL, Schwarzkopf DS, Maguire EA. Exploring the parahippocampal cortex response to high and low spatial frequency spaces. Neuroreport 2012;23:503–7. doi:10.1097/WNR.0b013e328353766a.

[137] Cant JS, Goodale MA. Attention to Form or Surface Properties Modulates Different Regions of Human Occipitotemporal Cortex. Cereb Cortex 2007;17:713–31. doi:10.1093/cercor/bhk022.

[138] Cant JS, Xu Y. Object Ensemble Processing in Human Anterior-Medial Ventral Visual Cortex. J Neurosci 2012;32:7685–700. doi:10.1523/JNEUROSCI.3325-11.2012.

[139] Walther DB, Caddigan E, Fei-Fei L, Beck DM. Natural scene categories revealed in distributed patterns of activity in the human brain. Journal of Neuroscience 2009;29:10573–81.

[140] Peelen MV, Fei-Fei L, Kastner S. Neural mechanisms of rapid natural scene categorization in human visual cortex. Nature 2009;460:94–7.

[141] Park S, Intraub H, Yi D-J, Widders D, Chun MM. Beyond the Edges of a View: Boundary Extension in Human Scene-Selective Visual Cortex. Neuron 2007;54:335–42. doi:10.1016/j.neuron.2007.04.006.

[142] Chadwick MJ, Mullally SL, Maguire EA. The hippocampus extrapolates beyond the view in scenes: An fMRI study of boundary extension. Cortex 2013;49:2067–79. doi:10.1016/j.cortex.2012.11.010.

[143] MacEvoy SP, Epstein RA. Decoding the representation of multiple simultaneous objects in human occipitotemporal cortex. Current Biology 2009;19:943–7.

[144] Troiani V, Stigliani A, Smith ME, Epstein RA. Multiple Object Properties Drive Scene-Selective Regions. Cereb Cortex 2012. doi:10.1093/cercor/bhs364.

[145] Epstein RA, Ward EJ. How Reliable Are Visual Context Effects in the Parahippocampal Place Area? Cereb Cortex 2010;20:294–303. doi:10.1093/cercor/bhp099.

[146] Aminoff EM, Kveraga K, Bar M. The role of the parahippocampal cortex in cognition. Trends in Cognitive Sciences 2013;17:379–90. doi:10.1016/j.tics.2013.06.009.

[147] Bar M, Aminoff E. Cortical analysis of visual context. Neuron 2003;38:347–58.

[148] Stevens WD, Kahn I, Wig GS, Schacter DL. Hemispheric Asymmetry of Visual Scene Processing in the Human Brain: Evidence from Repetition Priming and Intrinsic Activity. Cereb Cortex 2012;22:1935–49. doi:10.1093/cercor/bhr273.

[149] Koutstaal W, Wagner AD, Rotte M, Maril A, Buckner RL, Schacter DL. Perceptual specificity in visual object priming: functional magnetic resonance imaging evidence for a laterality difference in fusiform cortex. Neuropsychologia 2001;39:184–99.

[150] Xu Y, Turk-Browne NB, Chun MM. Dissociating Task Performance from fMRI Repetition Attenuation in Ventral Visual Cortex. J Neurosci 2007;27:5981–5. doi:10.1523/JNEUROSCI.5527-06.2007.

[151] Nasr S, Devaney KJ, Tootell RBH. Spatial encoding and underlying circuitry in scene-selective cortex. NeuroImage 2013;83:892–900. doi:10.1016/j.neuroimage.2013.07.030.

[152] Baldassano C, Beck DM, Fei-Fei L. Differential connectivity within the Parahippocampal Place Area. NeuroImage 2013;75:228–37. doi:10.1016/j.neuroimage.2013.02.073.

[153] Biederman I. Perceiving real-world scenes. Science (New York, NY) 1972;177:77–80.

[154] Oliva A, Schyns PG. Diagnostic colors mediate scene recognition. Cognitive Psychology 2000;41:176–210. doi:10.1006/cogp.1999.0728.

[155] Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 2001;42:145–75.

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

[157] Vogel J, Schiele B. Semantic Modeling of Natural Scenes for Content-Based Image Retrieval. International Journal of Computer Vision 2007;72:133–57.

[158] Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 1996;381:607–9. doi:10.1038/381607a0.

[159] Simoncelli EP, Olshausen BA. Natural image statistics and neural representation. Annu Rev Neurosci 2001;24:1193–216. doi:10.1146/annurev.neuro.24.1.1193.

[160] Barlow H. Possible principles underlying the transformations of sensory messages. In: Rosenblith W, editor. Sensory Communication, MIT Press; 1961, p. 217–34.

[161] Fiser J, Aslin RN. Unsupervised Statistical Learning of Higher-Order Spatial Structures from Visual Scenes. Psychological Science 2001;12:499–504. doi:10.1111/1467-9280.00392.

[162] Fiser J, Aslin RN. Encoding multielement scenes: statistical learning of visual feature hierarchies. J Exp Psychol Gen 2005;134:521–37. doi:10.1037/0096-3445.134.4.521.

[163] Elder JH, Goldberg RM. Ecological statistics of Gestalt laws for the perceptual organization of contours. J Vis 2002;2:324–53. doi:10.1167/2.4.5.

[164] Geisler WS, Perry JS, Super BJ, Gallogly DP. Edge co-occurrence in natural images predicts contour grouping performance. Vision Res 2001;41:711–24.

[165] Samonds JM, Potetz BR, Lee TS. Cooperative and competitive interactions facilitate stereo computations in macaque primary visual cortex. J Neurosci 2009;29:15780–95. doi:10.1523/JNEUROSCI.2305-09.2009.

[166] Turk-Browne NB, Jungé J, Scholl BJ. The automaticity of visual statistical learning. J Exp Psychol Gen 2005;134:552–64. doi:10.1037/0096-3445.134.4.552.

[167] Bar M, Aminoff E. Cortical analysis of visual context. Neuron 2003;38:347–58.

[168] Schapiro AC, Rogers TT, Cordova NI, Turk-Browne NB, Botvinick MM. Neural representations of events arise from temporal community structure. Nat Neurosci 2013;16:486–92. doi:10.1038/nn.3331.

[169] Zadra JR, Proffitt DR. Implicit Associations Have a Circadian Rhythm. PLoS ONE 2014;9:e110149. doi:10.1371/journal.pone.0110149.

[170] Szpiro SFA, Carrasco M. Exogenous Attention Enables Perceptual Learning. Psychol Sci 2015;26:1854–62. doi:10.1177/0956797615598976.

[171] Turk-Browne NB, Simon MG, Sederberg PB. Scene Representations in Parahippocampal Cortex Depend on Temporal Context. J Neurosci 2012;32:7202–7. doi:10.1523/JNEUROSCI.0942-12.2012.

[172] Fiser J, Berkes P, Orbán G, Lengyel M. Statistically optimal perception and learning: from behavior to neural representations. Trends Cogn Sci 2010;14:119–30. doi:10.1016/j.tics.2010.01.003.

[173] Brady TF, Oliva A. Statistical Learning Using Real-World Scenes Extracting Categorical Regularities Without Conscious Intent. Psychological Science 2008;19:678–85. doi:10.1111/j.1467-9280.2008.02142.x.

[174] Torralba A, Oliva A, Castelhano MS, Henderson JM. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol Rev 2006;113:766–86. doi:10.1037/0033-295X.113.4.766.

[175] Neider MB, Zelinsky GJ. Scene context guides eye movements during visual search. Vision Res 2006;46:614–21. doi:10.1016/j.visres.2005.08.025.

[176] Chun MM, Jiang Y. Contextual cueing: implicit learning and memory of visual context guides spatial attention. Cogn Psychol 1998;36:28–71. doi:10.1006/cogp.1998.0681.

[177] Mayr U. Spatial attention and implicit sequence learning: evidence for independent learning of spatial and nonspatial sequences. J Exp Psychol Learn Mem Cogn 1996;22:350–64.

[178] Berry D, Dienes Z. Towards a Characterization of Implicit Learning. In: Berry DC, Dienes Z, editors. Implicit Learning: Theoretical and Empirical Issues, Lawrence Erlbaum Associates; 1993, p. 1–18.

[179] Dienes Z, Berry D. Implicit learning: Below the subjective threshold. Psychonomic Bulletin & Review 1997;4:3–23. doi:10.3758/BF03210769.

[180] Saffran JR, Johnson EK, Aslin RN, Newport EL. Statistical learning of tone sequences by human infants and adults. Cognition 1999;70:27–52.

[181] Summerfield JJ, Lepsien J, Gitelman DR, Mesulam MM, Nobre AC. Orienting attention based on long-term memory experience. Neuron 2006;49:905–16. doi:10.1016/j.neuron.2006.01.021.

[182] Brady TF, Chun MM. Spatial constraints on learning in visual search: modeling contextual cuing. J Exp Psychol Hum Percept Perform 2007;33:798–815. doi:10.1037/0096-1523.33.4.798.

[183] Leibo JZ, Cornebise J, Gómez S, Hassabis D. Approximate Hubel-Wiesel Modules and the Data Structures of Neural Computation. arXiv:151208457 [cs, Q-Bio] 2015.

[184] Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 1962;160:106–54.2.

[185] Serre T, Oliva A, Poggio T. A feedforward architecture accounts for rapid categorization. PNAS 2007;104:6424–9. doi:10.1073/pnas.0700622104.

[186] Zhao J, Turk-Browne NB. Incidental encoding of numerosity in visual long-term memory. Visual Cognition 2011;19:928–55. doi:10.1080/13506285.2011.598482.

[187] Ariely D. Seeing Sets: Representation by Statistical Properties. Psychological Science 2001;12:157–62. doi:10.1111/1467-9280.00327.

[188] Alvarez GA, Oliva A. The representation of simple ensemble visual features outside the focus of attention. Psychol Sci 2008;19:392–8. doi:10.1111/j.1467-9280.2008.02098.x.

[189] Goroshin R, Bruna J, Tompson J, Eigen D, LeCun Y. Unsupervised Learning of Spatiotemporally Coherent Metrics. arXiv:14126056 [cs] 2014.

[190] Fiser J, Aslin RN. Statistical learning of new visual feature combinations by infants. PNAS 2002;99:15822–6. doi:10.1073/pnas.232472899.

[191] Kirkham NZ, Slemmer JA, Johnson SP. Visual statistical learning in infancy: evidence for a domain general learning mechanism. Cognition 2002;83:B35–42. doi:10.1016/S0010-0277(02)00004-5.

[192] Kim R, Seitz A, Feenstra H, Shams L. Testing assumptions of statistical learning: is it long-term and implicit? Neurosci Lett 2009;461:145–9. doi:10.1016/j.neulet.2009.06.030.

[193] Zhao J, Al-Aidroos N, Turk-Browne NB. Attention is spontaneously biased toward regularities. Psychol Sci 2013;24:667–77. doi:10.1177/0956797612460407.

[194] Biederman I, Glass AL, Stacy EW. Searching for objects in real-world scences. J Exp Psychol 1973;97:22–7.

[195] Turk-Browne NB, Scholl BJ, Chun MM, Johnson MK. Neural Evidence of Statistical Learning: Efficient Detection of Visual Regularities Without Awareness. J Cogn Neurosci 2009;21:1934–45. doi:10.1162/jocn.2009.21131.

[196] Turk-Browne NB, Scholl BJ, Johnson MK, Chun MM. Implicit perceptual anticipation triggered by statistical learning. J Neurosci 2010;30:11177–87. doi:10.1523/JNEUROSCI.0858-10.2010.

[197] Bar M. Visual objects in context. Nat Rev Neurosci 2004;5:617–29. doi:10.1038/nrn1476.

[198] Aminoff E, Gronau N, Bar M. The Parahippocampal Cortex Mediates Spatial and Nonspatial Associations. Cerebral Cortex 2006;17:1493–503.

[199] Bar M, Aminoff E, Schacter DL. Scenes unseen: the parahippocampal cortex intrinsically subserves contextual associations, not scenes or places per se. J Neurosci 2008;28:8539–44. doi:10.1523/JNEUROSCI.0987-08.2008.

[200] Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. Science 2001;293:2425–30. doi:10.1126/science.1063736.

[201] Turk-Browne NB, Scholl BJ, Chun MM, Johnson MK. Neural Evidence of Statistical Learning: Efficient Detection of Visual Regularities Without Awareness. J Cogn Neurosci 2009;21:1934–45. doi:10.1162/jocn.2009.21131.

[202] Tambini A, Davachi L. Persistence of hippocampal multivoxel patterns into postencoding rest is related to memory. PNAS 2013;110:19591–6. doi:10.1073/pnas.1308499110.

[203] Tambini A, Ketz N, Davachi L. Enhanced brain correlations during rest are related to memory for recent experiences. Neuron 2010;65:280–90. doi:10.1016/j.neuron.2010.01.001.

[204] Westfall J, Yarkoni T. Statistically Controlling for Confounding Constructs Is Harder than You Think. PLOS ONE 2016;11:e0152719. doi:10.1371/journal.pone.0152719.

[205] McLaren DG, Ries ML, Xu G, Johnson SC. A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. Neuroimage 2012;61:1277–86. doi:10.1016/j.neuroimage.2012.03.068.

[206] Di Bernardi Luft C, Baker R, Bentham P, Kourtzi Z. Learning temporal statistics for sensory predictions in mild cognitive impairment. Neuropsychologia 2015;75:368–80. doi:10.1016/j.neuropsychologia.2015.06.002.

[207] Kravitz DJ, Peng CS, Baker CI. Real-world scene representations in high-level visual cortex: it's the spaces more than the places. J Neurosci 2011;31:7322–33. doi:10.1523/JNEUROSCI.4588-10.2011.

[208] Bastin J, Vidal JR, Bouvier S, Perrone-Bertolotti M, Bénis D, Kahane P, et al. Temporal components in the parahippocampal place area revealed by human intracerebral recordings. J Neurosci 2013;33:10123–31. doi:10.1523/JNEUROSCI.4646-12.2013.

[209] Aminoff EM, Tarr MJ. Associative Processing Is Inherent in Scene Perception. PLoS One 2015;10. doi:10.1371/journal.pone.0128840.

[210] Epstein RA. Parahippocampal and retrosplenial contributions to human spatial navigation. Trends Cogn Sci 2008;12:388–96. doi:10.1016/j.tics.2008.07.004.

[211] Marchette SA, Vass LK, Ryan J, Epstein RA. Outside Looking In: Landmark Generalization in the Human Navigational System. J Neurosci 2015;35:14896–908. doi:10.1523/JNEUROSCI.2270-15.2015.

[212] Baldassano C, Beck DM, Fei-Fei L. Differential connectivity within the Parahippocampal Place Area. Neuroimage 2013;75:228–37. doi:10.1016/j.neuroimage.2013.02.073.

[213] Khaligh-Razavi S-M, Kriegeskorte N. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLOS Comput Biol 2014;10:e1003915. doi:10.1371/journal.pcbi.1003915.

[214] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Object Detectors Emerge in Deep Scene CNNs. arXiv:14126856 [cs] 2014.

[215] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems 25, Curran Associates, Inc.; 2012, p. 1097–105.

[216] Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A. SUN database: Large-scale scene recognition from abbey to zoo, IEEE; 2010, p. 3485–92. doi:10.1109/CVPR.2010.5539970.

[217] Samonds JM, Potetz BR, Tyler CW, Lee TS. Recurrent connectivity can account for the dynamics of disparity processing in V1. J Neurosci 2013;33:2934–46. doi:10.1523/JNEUROSCI.2952-12.2013.

[218] Geisler WS, Perry JS, Super BJ, Gallogly DP. Edge co-occurrence in natural images predicts contour grouping performance. 2001.

[219] Kravitz DJ, Saleem KS, Baker CI, Mishkin M. A new neural framework for visuospatial processing. Nat Rev Neurosci 2011;12:217–30. doi:10.1038/nrn3008.