American Psychological Association

Journal of Experimental Psychology:
Human Perception and Performance

0096-1523/19/$12.00

2019, Vol. 45, No. 6, 715–728
http://dx.doi.org/10.1037/xhp0000626

Perceptual Dissociations Among Views of Objects, Scenes, and Reachable Spaces

Emilie L. Josephs and Talia Konkle
Harvard University

In everyday experience, we interact with objects and we navigate through space. Extensive research has revealed that these visual behaviors are mediated by separable object-based and scene-based processing mechanisms in the mind and brain. However, we also frequently view near-scale spaces, for example, when sitting at the breakfast table or preparing a meal. How should such spaces (operationalized here as "reachspaces"), which contain multiple objects but not enough space to navigate through, be considered in this dichotomy? Here, we used visual search to explore the possibility that reachspace views are perceptually distinctive from full-scale scene views as well as object views. In the first experiment, we found evidence for this dissociation. In the second experiment, we found that the perceptual differences between reachspaces and scenes were substantially larger than those between scene categories (e.g., kitchens vs. offices). Finally, we provide computational support for this perceptual dissociation: Deep neural network models also naturally separate reachspaces from both scenes and objects, suggesting that mid- to high-level features may underlie this dissociation. Taken together, these results demonstrate that our perceptual systems are sensitive to systematic visual feature differences that distinguish objects, reachspaces, and full-scale scene views. Broadly, these results raise the possibility that our visual system may use different perceptual primitives to support the perception of reachable and navigable views of the world.

Public Significance Statement
The present study suggests that views of near-scale space (e.g., kitchen counters, office desktops) look systematically different than views of far-scale space (e.g., full-scale kitchens, full-scale offices). These perceptual differences were substantially larger than those between scene categories (e.g., kitchens vs. offices).

Keywords: visual search, visual features, reachspaces, scenes, objects

As we behave in the world, there is a clear distinction between spatially compact elements of the environment that we can hold and spatially extended elements that we can move through, namely, between objects and scenes (Epstein, 2005; Henderson & Hollingworth, 1999). Extensive evidence from development, neuropsychology, and cognition research suggests that this fundamental division is reflected in our cognitive architecture, with different developmental trajectories, disorders, and processing demands associated with each (e.g., D. P. Carey, Dijkerman, Murphy, Goodale, & Milner, 2006; S. Carey & Xu, 2001; Epstein, Deyoe, Press, Rosen, & Kanwisher, 2001; Henderson & Hollingworth, 1999; Landis, Cummings, Benson, & Palmer, 1986; Spelke, 1990; Steeves et al., 2004). This distinction is also evident in the visual processing stream, in which distinct brain regions are sensitive to object-based and scene-based perceptual properties, and support object- and scene-related processing, respectively (e.g., Biederman, 1987; Epstein & Kanwisher, 1998; Greene & Oliva, 2009, 2010; Grill-Spector, Kourtzi, & Kanwisher, 2001).

However, real-world views of objects and scenes exist along a continuum: between the close-up view of a carrot and the far-scale view of the kitchen is the near-scale view of the countertop on which you prepared the food. Consider also the view of the desk as you type an e-mail, or a workbench as you solder a wire, or a place setting as you eat a meal. These views, which we here refer to as "reachspaces" (see Figure 1), are like scenes, in that they extend beyond the view and contain multiple objects. But unlike scenes, you do not navigate your body through them; instead, the interaction demands are more object-like in that they involve coordinated hand actions.

Such intermediate views have historically been treated as equivalent to navigable-scale scenes (Intraub, 2010) or have been acknowledged as an uneasy fit in the object–scene dichotomy and

Emilie L. Josephs and Talia Konkle, Department of Psychology, Harvard University.

We thank Bolei Zhou and Chen-Ping Yu for their assistance with the computational modeling components, Chen Ping for building and training the DNNs from which we extracted visual features, Roger Strong for his help with sample size simulation, and Sarah Cohen for her assistance with data collection. Stimuli and data for the behavioral experiments have been uploaded to an Open Science Framework repository (https://osf.io/7j6cx/; Josephs & Konkle, 2018).

Correspondence concerning this article should be addressed to Emilie L. Josephs, Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02143. E-mail: ejosephs@g.harvard.edu

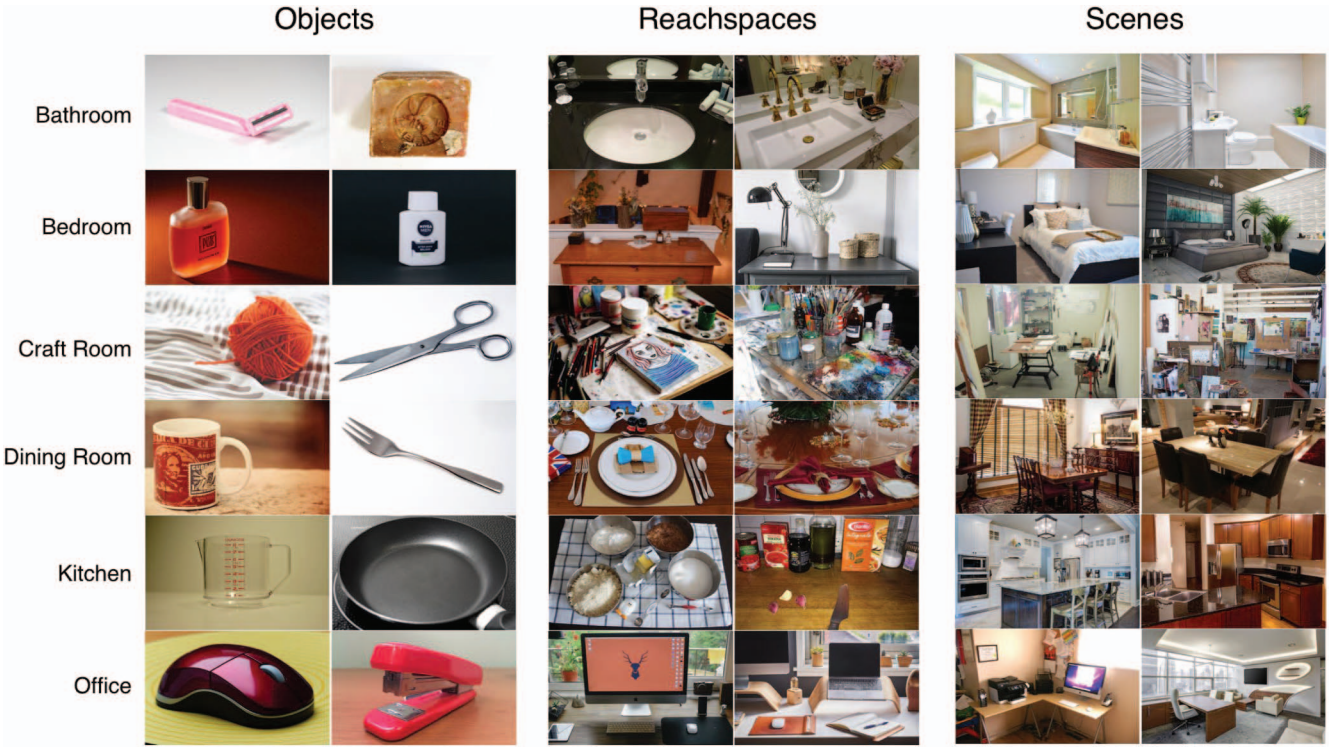

Figure 1. Examples of the object, reachspace, and scene stimuli. Object images consisted of single objects on their natural backgrounds. Reachspace images consisted of close-scale spaces, delineated by a horizontal surface and filled with objects, where everyday tasks are typically performed. Scene images consisted of full views of indoor rooms. Note that the photographs used in this figure are not the images used in the experiment, but rather are copyright-free images selected to closely match the experimental stimuli (see appendix for image attributions). For examples of the actual images used in the experiment, see our repository at osf.io/7j6cx. See the online article for the color version of this figure.

then consequently omitted from scene perception research (cf. Henderson & Hollingworth, 1999). However, this omission may have obscured important distinctions, as previous work suggests that the ways in which we act, remember, and deploy attention may differ for near and far space. For example, boundary extension, the anticipatory representation of the space beyond the edge of an image, is stronger for near space than far space (Bertamini, Jones, Spooner, & Hecht, 2005; Intraub, Bender, & Mangels, 1992). Additionally, attention can be impaired for spaces near the body but not farther way, and vice versa, in patients with hemispatial neglect (Cowey, Small, & Ellis, 1994; Halligan & Marshall, 1991). Finally, evidence from the visuomotor literature highlights that the distinction between what is in and out of reach is a prominent one, with near-space coding evident extensively across dorsal stream regions (e.g., Gallivan, Cavina-Pratesi, & Culham, 2009; Maravita & Iriki, 2004). Thus, there is both behavioral and neural evidence that there may be important differences in the processing of near and far space.

In the present study, we asked whether these functional differences between scales of space run alongside *perceptual differences* in their visual appearance. Specifically, we tested whether there are systematic perceptual differences between reachspaces and scenes. On one hand, both reachspaces and scenes depict extended surfaces with multiple objects, and thus may rely on common perceptual features. Consistent with this reasoning, some studies have used views of reachable spaces to highlight mechanisms of scene perception, with the assumption that near- and far-scale views can be used interchangeably (e.g., Epstein, Graham, & Downing, 2003; Võ & Wolfe, 2013). On the other hand, reachspaces contain small objects, whereas full-scene views are dominated by large objects, and small and large objects have dissociable perceptual features (Long, Konkle, Cohen, & Alvarez, 2016). There may also be differences in their global layout features, as previous computational models have leveraged natural image statistics to estimate the depicted depth in an image along the full object–scene continuum (e.g., Torralba & Oliva, 2002, 2003). These results raise the possibility that there are different image statistics for near- and far-scale space, which human perceptual systems may be sensitive to, enabling views of reachspaces to dissociate from views of scenes in perception.

We also examined whether reachspace views perceptually dissociate from object views. Although objects are bounded entities, images of objects on a naturalistic background must necessarily depict some space, leading to some ambiguity with reachspace views. However, it is possible to operationalize the differences between these two kinds of views in terms of their implied viewing distances (cf. Intraub, 2010, 2012). Object views gives a sense of a very close viewing distance (~8–12 in. away) and feature a

central object on a homogeneous background, cropped so that it fills much of the image, with minimal to no "layout" edges, for example, where a counter meets a back wall or a corner. In contrast, reachspace views convey a sense of the space about of 3 to 4 ft. from the viewer, have salient 3D layout features (e.g., a flat horizontal surface and a back wall), and contain an array of contextually related objects. Object views are scaled such that a hand would fill the frame, whereas reachspace views are scaled such that both arms would fit in the space. Thus, object and reachspace views are sampled from systematically different views of the environment, and as such, they may have different visual feature properties.

To examine whether reachspaces are perceptually distinct from scenes and objects, we employed both behavioral and computational methods. First, we used a visual search paradigm, in which the speed of search depends on how visually distinctive the target is from the distractors (Duncan & Humphreys, 1989; Wolfe & Horowitz, 2017). Across several experiments, we found that reachspaces systematically dissociate from both full-scale scenes and singleton objects in search displays, with an effect that is substantially larger than the perceptual difference between semantic categories of scenes. Second, examining deep convolutional neural networks trained to perform either object or scene recognition, we found that both kinds of networks naturally distinguish reachspace views from both objects and scenes in middle and later layers. This result provides computational support for the existence of a distinctive visual representation of reachspaces and begins to address questions about the nature of these feature distinctions. Broadly, these results raise the possibility that there may be separate perceptual processing mechanisms for reachable and navigable space.

Experiment 1: A Three-Way Dissociation in Visual Search Performance

To explore the possibility that reachspaces perceptually dissociate from objects and scenes, we used a visual search paradigm (e.g., following Cohen, Alvarez, Nakayama, & Konkle, 2017; Long, Störmer & Alvarez, 2017). Under the logic of a visual search task, targets that are different from distractors will stand out and will be faster to find in a search array than targets that are similar (Duncan & Humphreys, 1989). Thus, for example, if reachspaces are perceptually distinct from scenes, then it should be easier to find a reachspace among scenes than among other reachspaces. Visual search speeds are strongly influenced by visual similarity (e.g., leveraging feature differences between line orientations, curvature, and shape) and are largely unaffected by nonvisual (semantic) information in the displayed items (see Wolfe & Horowitz, 2017). Thus, any differences in visual search times between objects, reachspaces, and scenes would provide evidence for a dissociation at the level of visual perception. Experiment 1 tests this hypothesis on two image sets: images in Experiment 1a are matched in luminance and contrast, whereas images in Experiment 1b are additionally matched in global spatial frequency content. This image set manipulation enables us to test for visual differences over and above relatively primitive global image statistics.

Method

Participants. Forty-four participants were enrolled ($N = 22$ each in Experiments 1a and 1b). These sample sizes were estimated to provide 80% power using a simulation method (see Appendix for details). Demographic information was not recorded from individual participants, but all participants were between the ages of 18 and 35 years, had normal or corrected to normal vision, and were recruited from a participant population that consisted of 65% women. All participants gave informed consent and were compensated with $10 or class credit for their participation. All procedures were approved by the Harvard University Human Subjects Institutional Review Board.

Stimuli. The stimulus set consisted of views of objects, reachspaces, and scenes (examples in Figure 1). Each of these image scales contained 12 images from each of six semantic categories (bathroom, bedroom, craft room, dining room, kitchen, office), yielding 72 images per scale. Object images depicted close-scale views (within 8–12 in. from the object) of single objects on their natural background, for example, a close-up view of a sponge with a small amount of granite countertop visible beyond it. Reachspace images depicted near-scale spaces that were approximately as deep as arm's reach (3–4 ft.), consisting of multiple small objects arrayed on a horizontal surface, for example, a knife, cutting board, and an onion arrayed on kitchen counter. Scene images depicted full views of the interior of rooms, for example, a view of a home office. Images were collected from Google Images under fair use and were scaled to a resolution of 800 × 1280 pixels. All stimuli are available for download on the Open Science Framework (https://osf.io/7j6cx).

Images were controlled using the SHINE toolbox (Willenbockel et al., 2010) to be matched in their average luminance (Experiment 1a), and in both average luminance and in spatial frequency (Experiment 1b). Examples are shown in Figure 2B. Images for Experiment 1a were luminance matched using the lumMatch function run with the default settings, and images for Experiment 2 were spatial-frequency matched using the specMatch function, then luminance matched using the histMatch function, both with default parameters.

Design. Participants searched for a single target in an array of distractors (Figure 2A). Each trial started with a fixation cross in the center of the screen for 500 ms. Next, a preview of the target appeared in the center of the screen for 500 ms. Following a second 500-ms fixation screen, the search display appeared, consisting of six images arranged in a circle to be equidistant from the center of the screen. No images were placed on the vertical midline. One of these images was always the target. Participants pressed the spacebar when they found the target, and then all images in the display were replaced by Xs. Participants clicked on the X corresponding to the target location. For correct responses, the next trial would start after a variable time interval lasting between 500 ms and 1,000 ms, but for incorrect responses, participants received a feedback message for 3 s before the next trial began.

The experiment was a 3 × 3 design: On a given trial, a target could be an object, a reachspace, or a scene, and could be displayed among distractors that were all objects, all reachspaces, or all scenes, leading to a fully-crossed design with a total of nine conditions. There were 50 trials in each of the nine conditions. On

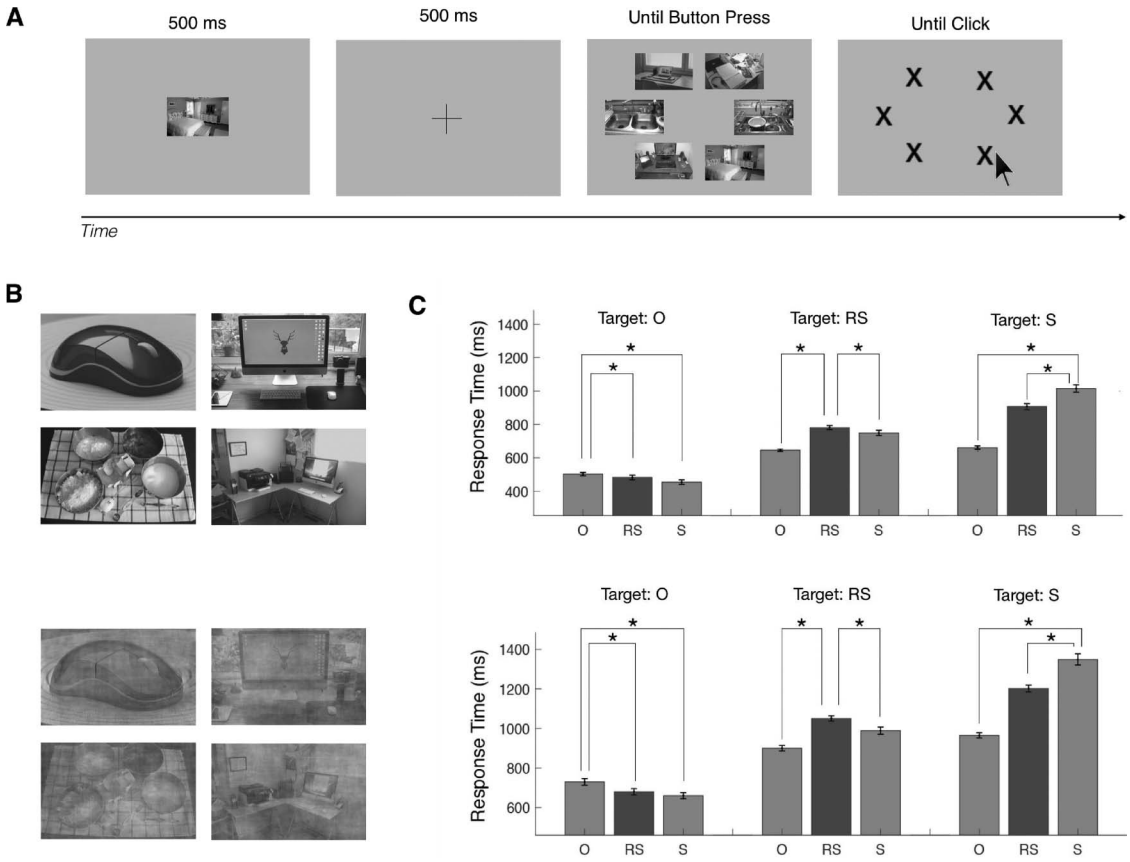

Figure 2. Trial design, stimuli and results for Experiment 1. A. Example time course of a single trial. A target was presented for 500 ms, followed by a 500 ms blank, followed by the search display. The search display contained 6 items, one of which was the target. Participants pressed the space bar when they found the target, replacing images by Xs, then clicked the location of the target with the mouse. B. The stimuli for Experiment 1a (top) and 1b (bottom). Images in Experiment 1a were matched in mean luminance. Images in Experiment 1b were matched in both luminance and spatial frequency. C. Reaction time results for Experiment 1a (top) and 1b (bottom). Reaction times is plotted for each target-distractor combination, with the first set of bars reflecting Object targets, the second set of bars reflecting Reachspace targets, and the third set of bars reflecting Scene targets. Stars mark significant differences. Error bars show within-subject standard error of the mean (Morey, 2008). Note that the photographs used in this figure are not the images used in the experiment, but rather are copyright-free images selected to closely match the experimental stimuli (see appendix for image attributions). For examples of the actual images used in the experiment, see our repository at osf.io/7j6cx.

every trial, targets and distractors were randomly selected from the appropriate image scale, with the constraint that none of them could have the same semantic category as the target. Thus, for example, a kitchen reachspace target could appear among scene distractors that included bathrooms, bedroom, craft rooms, dining rooms, or offices, but not kitchens. The experiment began with 10 practice trials (practice images were drawn from the same set used in the experimental trials), followed by nine blocks of 50 experimental trials.

Apparatus. Experiments were run on a 24-in. iMac running OS 5.10.8 in MATLAB 7.10.0 (The MathWorks, Natick, MA) using the Psychophysics Toolbox (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). The monitor was set to a spatial resolution of 1,920 × 1,200 pixels and a refresh rate of 60 Hz. Observers were seated approximately 57 cm from the monitor, so 1 cm on the screen subtended 1° of visual angle. All images were shown at a size of 8° × 5° visual angle (300 × 188 pixels). Responses were recorded on a standard Apple Keyboard.

Data analysis. Reaction time (RT) data were trimmed to remove outliers using the following procedure. First, incorrect trials and trials with RTs less than 200 ms were excluded. Mean RT was then calculated for each condition and each participant, and RTs that fell more than three standard deviations away from the mean (for a given participant on a given condition) were discarded (Rousseeuw & Croux, 1993). RTs were log transformed prior to trimming to account for the fact that RT distributions are right-skewed (Palmer, Horowitz, Torralba, & Wolfe, 2011; Ratcliff, 1979). This procedure led to the exclusion of 1.2% of trials from Experiment 1a and 0.6% of trials from Experiment 1b. Individual participants who lost more than 15% of their trials to this trimming were replaced. One subject was dropped from Experiment 1a for this reason and replaced, and no subjects were

dropped from Experiment 1b. Planned pairwise one-tailed t tests were used to assess statistical significance. Effect size was calculated using the classical Cohen's d: the difference between the means of the conditions divided by the pooled variance.

Results and Discussion

Figure 2C shows the results of the visual search experiments using luminance-matched images (Experiment 1a) and spatial-frequency-matched images (Experiment 1b). To interpret the pattern of results across the nine conditions (3 target view types × 3 distractor view types), we examined the conditions in pairs: for example, if objects are perceptually distinct from scenes, they should be found faster in displays of scene distractors than in displays of object distractors. Indeed, as expected, objects and scenes dissociate from each other: Object targets were found more quickly among scenes than among other objects (Experiment 1a, $t[21] = 4.59$, $p < .001$, $d = 0.52$; Experiment 1b, $t[21] = 4.20$, $p < .001$, $d = 0.48$). In the complementary comparison, scenes targets were likewise found more quickly among objects than among other scenes (Experiment 1a, $t[21] = 13.93$, $p < .001$, $d = 2.33$; Experiment 1b, $t[21] = 12.37$, $p < .001$, $d = 1.98$).

In the critical comparisons, we found that reachspaces dissociated from scenes. That is, reachspace targets were found faster among scenes than among other reachspaces (Experiment 1a, $t[21] = 2.09$, $p = .024$, $d = 0.22$; Experiment 1b, $t[21] = 3.04$, $p = .003$, $d = 0.34$), and likewise, scenes targets were found faster among reachspaces than other scenes (Experiment 1a, $t[21] = 3.92$, $p < .001$, $d = 0.62$; Experiment 1b, $t[21] = 5.64$, $p < .001$, $d = 0.74$). Additionally, we found that reachspaces dissociated from objects: Reachspace targets were found more quickly among objects than among other reachspaces (Experiment 1a, $t[21] = 8.52$, $p < .001$, $d = 1.02$; Experiment 1b, $t[21] = 7.51$, $p < .001$, $d = 0.92$), and corresponding object targets were found more quickly among reachspaces than among other objects (Experiment 1a, $t[21] = 1.74$, $p = .048$, $d = 0.21$; Experiment 1b, $t[21] = 3.18$, $p = .002$, $d = 0.35$).

These results, replicated across two experiments, provide evidence that reachspaces are perceptually different from both full-scale scenes and singleton objects. Given that both reachspaces and scenes have a spatial layout and consist of multiple objects, it could have been the case that they are visually encoded using the same perceptual features. However, the visual search data instead suggest that the perceptual content of views of reachable space is systematically different from the perceptual content of full-scale scenes. Furthermore, these behavioral dissociations persist in spatial-frequency matched images (Experiment 1b), indicating that differences in global spatial frequency or luminance content are not solely responsible for distinguishing reachspaces from objects and scenes. Finally, these data do not require that objects, reachspaces, and scenes be separate categories in the mind a priori; it is possible that participants could perform this task using on-the-fly categories developed in the context of this particular visual search design. However, critically, the main conclusion that there are perceptual feature differences between these three scales does not depend on the exact strategy used in performing the task.

In addition, broader trends in the data shed some light on the nature of this three-way dissociation, suggesting that reachspaces are perceptually intermediate. That is, the RT differences imply that objects and scenes are the most perceptually dissimilar and that reachspaces come somewhere in between: Scenes were found more quickly among objects than among reachspaces (Experiment 1a, $t[21] = 13.02$, $p < .001$, $d = 1.68$; Experiment 1b, $t[21] = 11.32$, $p < .001$, $d = 1.44$; post hoc paired one-sided t test). Likewise, objects were found more quickly among scenes than among reachspaces (Experiment 1a, $t[21] = 5.55$, $p < .001$, $d = 0.35$; Experiment 1b did not reach significance, $t[21] = 1.27$, $p = .106$, $d = 0.14$). Additionally, we find further evidence that reachspaces are perceptually intermediate when considering overall search times by target: Object targets, overall, were found faster than reachspaces, which were found faster than scenes (Experiment 1a main effects: object targets = 480 ms; reachspaces = 726 ms; scenes = 861 ms; 3 × 2 ANOVA main effect of scale, $F[2, 194] = 111.38$, $p < .001$; Experiment 1b main effects: object targets = 689 ms; reachspaces = 979 ms; scenes = 1,172 ms; 3 × 2 ANOVA main effect of scale, $F[2, 194] = 119.36$, $p < .001$). Taken together, these visual search dissociations demonstrate that reachspaces are perceptually distinguishable from, and intermediate to, objects and scenes.

Experiment 2: Pitting the Effect of Image Scale Against Scene Category

How substantial are the perceptual differences between reachspaces and full-scale scenes? Will *any* meaningful distinction among visual environments give rise to similar visual search effects, or do the present dissociations constitute a particularly sizable difference? To put the perceptual dissociation between reachspaces and scenes into context, we compared it with an alternative distinction that is critical for scene processing: semantic category. It is clear that the visual system is sensitive to perceptual features that help distinguish among scene categories: Within a brief glance, observers can readily identify the semantic category of a scene, for example, whether it is a kitchen or an office (Potter, 1975, 1976). The aim of Experiment 2 was to assess the magnitude of scene-category effects on visual search time and compare this with the magnitude of the reachspace-scene effect. Specifically, we tested whether searching for a scene target was faster when the distractors differed by scene category, by scale of space, or both (Figure 3A).

Method

Participants. Twenty-seven participants were enrolled in Experiment 2. This sample size was estimated to provide 80% power according to the same simulation used for Experiment 1 (see Appendix). As before, participants were between the ages of 18 and 35 years and had normal or corrected-to-normal vision. All participants gave informed consent, were recruited from the Harvard Psychology Department participant pool, and were compensated with $10 or class credit for their participation. All procedures were approved by the Harvard University Human Subjects Institutional Review Board.

Stimuli. The stimulus set consisted of 72 images of scenes from six categories and 72 images of reachspaces from six categories (this constituted all the scene and reachspace images from Experiment 1a). These images were matched in average luminance.

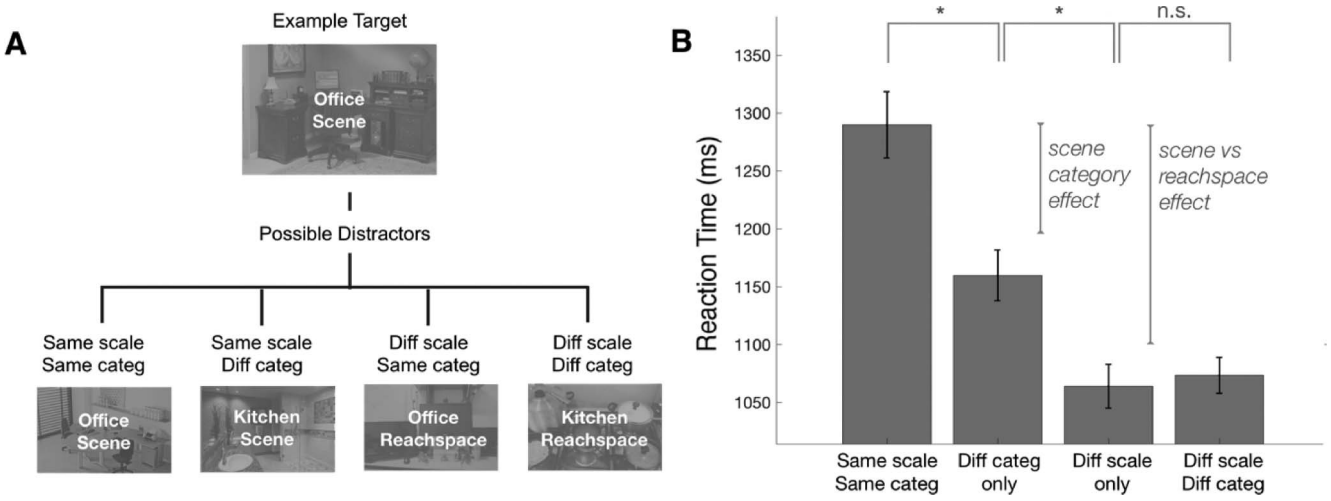

Figure 3. Trial design and results for Experiment 2. (A) The four conditions in Experiment 2. The target in each trial was always a scene image, drawn from one of six semantic categories. These targets were shown among distractors that could be either of the same or different image scale and the same or different semantic category. (B) Reaction times are plotted for the four conditions: same view type same category; same view type different category; different view type different category; different view type different category. Error bars represent within-subject standard error of the mean (Morey, 2008). See appendix for image attributions.

Design. Participants searched for a target among distractors with the same trial timing and display parameters as in Experiment 1 but with different target-distractor conditions (Figure 3A). The target in this experiment was always a scene image. The distractors could be either reachspaces or scenes, drawn from either the same or different semantic category as the target scene. There were 30 trials in each of these four conditions (same scale/same category; different scale/same category; same scale/different category; different scale/different category). To counterbalance scene category, scene targets were drawn equally from the six semantic categories, appearing five times each in each condition. For conditions in which distractors were from a different semantic category than the target, these were balanced such that each target category appeared among each of the other five distractor categories exactly once. Trials were split into two blocks of 60 trials, with 11 practice trials at the beginning. Trial order was randomized over the experiment.

Data analysis. The data trimming procedure was the same as in Experiment 1. This procedure led to the exclusion of 2.0% of trials from Experiment 2. One subject was dropped from Experiment 2 for losing more than 15% of their trials to trimming, and was replaced. Planned pairwise one-tailed t tests were used to assess statistical significance, and Cohen's d was used to estimate effect size as described in Experiment 1.

Results and Discussion

Average RTs for Experiment 2 are shown in Figure 3B. Search was slowest when the target scene matched the distractors in both category and scale (e.g., a bathroom scene among other bathroom scenes; 1,290 ms, $SEM = 29$ ms), establishing a baseline condition from which to compare the other conditions. When distractors differed from the target in their semantic category alone (e.g., an office scene among dining room scenes), search was significantly faster than the baseline (130 ms faster, $t[26] = 3.51$ $p = .001$, $d = 0.57$). This result confirms that different scene categories are perceptually dissociable. Search was also significantly faster than baseline when the distractors differed in scale (e.g., an office scene among office reachspaces: 225 ms faster than baseline, $t[26] = 6.42$, $p < .001$, $d = 1.02$). This result confirms that reachspaces are perceptually dissociable from scenes. Crucially, the image scale effect was much larger than the semantic category effect (130 ms vs. 225 ms; post hoc paired one-sided t test, $t[26] = 3.15$, $p = .002$, $d = 0.51$). Finally, there was no additional speed to be gained when distractors differed in both image scale and scene category (9 ms difference between reachspace distractors of the same vs. different semantic category, $t[26] = 0.51$, $p = .31$, $d = 0.05$).

These results reveal that the perceptual distinction between reachspaces and scenes has a substantially larger impact on visual search behavior than perceptual distinctions between semantic categories. Further, these data serve as a replication of one of the key results in Experiment 1: Reachspaces are not "just scenes" when it comes to visual search behavior.

Visual Feature Analysis: Three-Way Dissociation in Computational Image Features

To this point, we have used RT-based behavioral tasks to infer that our perceptual systems are sensitive to systematic differences among different views of space. Such perceptual differences must arise from different image-computable visual statistics present in each of the views. Thus, we next turned to a computational modeling approach to bolster this claim with a computational existence proof of such feature differences and to provide some insight into the possible nature of these visual feature distinctions. Specifically, we used deep convolutional neural networks (DNNs), which can be treated as sophisticated pattern extractors, to measure visual feature differences among the views. Furthermore, because these models measure purely visual information—divorced from

semantic interpretation, prior information, or expectations—they provide a strong test of the hypothesis that the three scales of space explored here differ in their visual content.

DNNs are currently the state of the art for object and scene recognition by computer vision systems. Through extensive exposure to natural images, these models learn to detect particular visual features from an image in order to perform a specified task (e.g., object categorization, scene categorization). Critically, feature detector neurons are arranged in hierarchical layers, where early layers detect simpler image statistics and deeper layers detect increasingly complex features, based on weighted combinations of the features in the previous layer. Currently, the features learned by DNNs are the best model for predicting biological feature tuning along the visual processing stream, outperforming categorical models and other visual feature models (Cadieu et al., 2014; Cichy, Khosla, Pantazis, & Oliva, 2017; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014), and can account for some structure in behavioral judgments of objects similarity (Jozwik, Kriegeskorte, Storrs, & Mur, 2017; King, Groen, Steel, Kravitz, & Baker, 2018). This correspondence between models, brains, and behavior is especially noteworthy because these DNN models were not optimized to predict brains or behavioral similarity judgments, suggesting that artificial and biological systems have arrived at similar solutions of what visual features are useful for distinguishing among categories, even if learned in very different ways. Thus, deep neural networks provide a powerful tool for exploring visual feature dissociations using an artificial visual system (e.g., Bonner & Epstein, 2018; Groen et al., 2018).

The logic of our approach was to compute deep net feature responses to the images in the stimulus set and then test whether reachspaces dissociate from objects and scene images (Figure 4). We considered two different pretrained neural networks. The first model was a seven-layer AlexNet architecture trained to perform object recognition (Krizhevsky, Sutskever, & Hinton, 2012), henceforth, "ObjectNet," and the second model had the same architecture but was trained to perform scene recognition ("SceneNet"). By comparing these two networks, we can examine whether any distinctions between objects, reachspaces, and scenes depend on features learned in service of object-category or scene-category distinctions. For each layer in each model, we calculated the response of each feature detector to each image in the set. By comparing how well objects, reachspaces, and scenes dissociate in each layer, we can infer something about the complexity of any features that distinguish these classes. Finally, we repeated this process for original full-colored stimuli as well as the controlled stimuli of Experiment 1b. By comparing original and controlled image sets, we can verify that any dissociations persist across changes in low-level features, as they did in human visual search behavior.

Method

Deep neural networks. For each DNN, we used an AlexNet architecture (Krizhevsky et al., 2012). Images were input at size $224 \times 224 \times 3$ pixels. Layer 1 was a convolutional layer with 64 kernels of size $11 \times 11 \times 3$, with a stride of 4 pixels and padding of 2 pixels. Layer 2 was a convolutional layer with 192 kernels of size $5 \times 5 \times 64$, with a stride of 1 pixels and padding of 2 pixels. Layer 3 was a convolutional layer with 384 kernels of size $3 \times 3 \times 192$, with a stride of 1 pixel and padding of 1 pixel. Layer 4 was a convolutional layer with 256 kernels of size $3 \times 3 \times 384$, with a stride of 1 pixel and padding of 1 pixel. Layer 5 was a convolutional layer with 256 kernels of size $3 \times 3 \times 256$, with a stride of 1 pixel and padding of 1 pixel. Layer 6 and 7 were fully connected layers of 4,096 neurons each.

One instantiation of this architecture was trained to do 1,000-way object categorization on the ImageNet database (Russakovsky et al., 2015). A second instantiation of this architecture was trained to do 205-way scene categorization using the Places database (Zhou, Lapedriza, Xiao, Torralba, & Oliva, 2014). Both networks were built and trained using in-lab software.

The standard AlexNet training regime was adopted using a public code package[1] that was optimized for multithreaded DNN training in Torch7. Specifically, stochastic gradient descent optimization was used with 0.9 momentum, an initial learning rate of 0.02, and weight decay of 0.0005. Both the learning rate and the weight decay follow a predefined decreasing schedule using a minibatch size of 128, with 10,000 minibatches per epoch over a total of 55 training epochs. Standard data augmentation such as random horizontal flips and random 224×224 crops were performed during training.

Stimuli. The image set was the same as in Experiment 1, cropped and resized to 224×224 pixels to match the expected input size of Layer 1. Neural net activations were recorded for full-color versions of the images ("original") as well as luminance and spatial frequency-controlled versions ("controlled").

Feature extraction. After training on either ImageNet or Places205, DNN weights were frozen, and the activations to our stimulus images were measured. Image features were extracted separately from each layer of each network using the following procedure. For convolutional layers, feature vectors for each image were extracted before normalization and response pooling operations, and calculated as the summed total activation of each kernel over the whole image (i.e., summing over the neurons). For fully connected layers, the feature vector was simply the activations of each unit (no summing over the units required).

Analysis. First, to visualize the degree of similarity among views of different scales, we used multidimensional scaling, which projects the high-dimensional feature space into two dimensions, such that images that have more similar feature vectors are located closer together in space. Differences among images were computed using the Euclidean distance between feature vectors. Nonmetric MDS was then performed over this distance matrix, projecting the data into 2D space for visualization.

To quantify this similarity, we used k-means clustering to group the images into three different categories based on their feature vectors extracted for each layer ($k = 3$, squared Euclidian distance metric, 100 replicates). We evaluated how well the clustering solution matched the ground truth grouping of images by scale using the Rand index. This index considers the set of pairwise comparisons between each of the images. Two images grouped together in both the clustering solution and the ground truth are counted as a correct pairing, and two items assigned to *different* groups in both the clustering solution and the ground truth are counted as a correct pairing. Grouping accuracy was then computed as the number of correct pairings divided by the total number of pairs, and multiplied by 100 to yield the grouping accuracy

[1] see https://github.com/soumith/imagenet-multiGPU.torch.

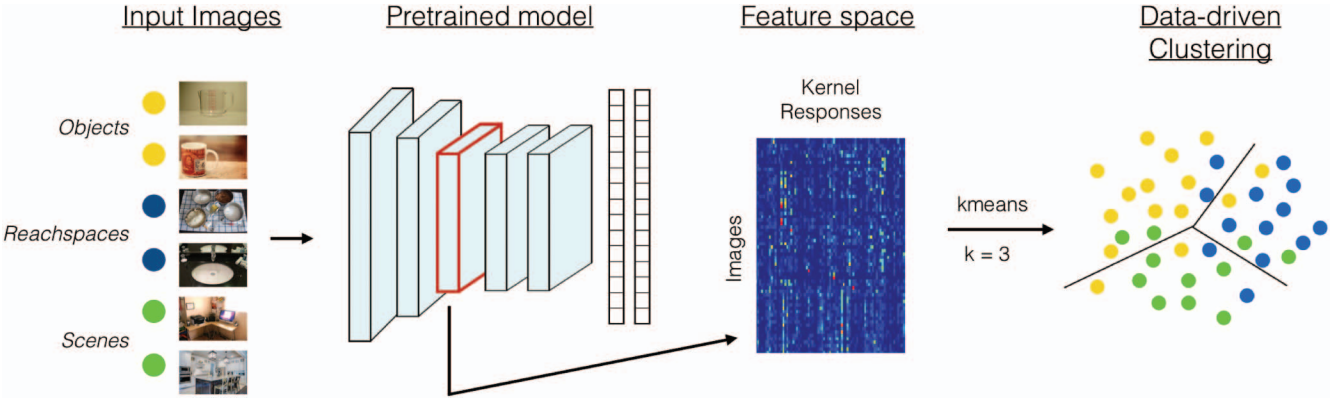

Figure 4. Extracting deep neural network features. We assessed whether data-driven clustering performed over the features could recapitulate the division between the three scales of space. Feature were extracted for each layer of a pre-trained deep neural network. For a given layer, kernel activations were extracted for every image in the set, then submitted to k-means clustering with k = 3, where images were assigned to a cluster based on their feature signatures. We then measured the correspondence between the clustering solution and the ground-truth grouping of the images by scale of space. See appendix for image attributions. See the online article for the color version of this figure.

percentage. This outcome measure takes a value of 1 if the clustering solution perfectly agrees with the ground truth but does not have a clear chance value. Thus, to estimate chance, simulations were run in which the grouping accuracy was computed when the object-reachspace-scene labels were randomly shuffled over the image set. Chance was set as the mean grouping accuracy percentage over 1,000 simulations. Finally, given that k-means clustering is a stochastic process, final clustering solutions depend in part on the locations of the randomly seeded initial centroids, and the final Rand index is not a stable number. To converge on a more stable estimate of the grouping accuracy, the clustering and scoring process was performed 100 times, and the average Rand index score is reported.

We also assessed the discriminability of the images using a naïve Bayes classifier with a leave-one-out cross-validation training scheme. Features with zero variance were dropped prior to fitting the model, and feature distributions were modeled using a kernel density estimator. The classifier predicted whether a held-out image was an object, reachspace, or scene based on the feature activations, and classifiers were fit separately for each layer of ObjectNet and SceneNet, for both original and controlled images. We additionally performed several auxiliary analyses to test the discriminability of other divisions in the image set. All auxiliary analysis were conducted using the naïve Bayes classifier and a leave-one-out cross-validation scheme.

Results and Discussion

First, we visualized the differences in DNN responses to objects, reachspaces, and scene images (see Figure 5). Multidimensional scaling was used to project the high-dimensional feature space captured in each DNN layer into a 2D plot, such that points that are more distant in the plot have more dissimilar feature activations. The figure shows feature spaces derived from a network trained to discriminate objects categories (Figure 5A) and a network trained to discriminate scene categories (Figure 5B) for both original and controlled variants of the image set. There are three main observations that are evident in this visualization. First, object, reachspace, and scene images do have different feature activations in both networks, becoming increasingly distinct in later layers. Second, reachspace images are distinguished from both and largely occupy intermediate positions relative to object and scene images. Third, these patterns also hold when images were equated in luminance and spatial frequency, particularly in later layers.

To quantify these observations, we used a data-driven clustering algorithm to divide the images into three clusters based on the similarity among their feature activations in each layer. Then, we calculated how well the data-driven clusters recovered the correct image classes (see Method, Figure 4). Note that this data-driven method simply finds the major joints in the visual feature space; it does not need any training or labels. Thus, any cases in which data-driven clusters correspond to the distinctions between image scales indicate that image scale is a major factor in the natural structure of the similarity space.

Results from the data-driven clustering analysis are shown in Figure 6 and confirm the patterns evident in the visualizations. That is, in ObjectNet, the different image scales were consistently assigned to separate clusters on the basis of their feature activations, in all layers beyond the lowest level feature representations of Layer 1 (grouping accuracy for each successive layer: 55%, 66%, 76%, 73%, 71%, 87% and 77%; simulated chance mean: 55%). The same results held when image features were extracted from SceneNet (grouping accuracy for each layer: 55%, 63%, 71%, 77%, 67%, 86% and 74%; simulated chance mean: 55%). Further, when considering the feature activations to the controlled images, all layers beyond the first two showed this same natural grouping (grouping accuracy ObjectNet: 56%, 54%, 65%, 62%, 67%, 67%, and 66%; SceneNet: 56%, 57%, 62%, 65%, 64%, 68%, and 67%; simulated chance mean: 55%).

To confirm that the results of the data-driven analysis generalized to other possible ways of analyzing the data, we also employed a

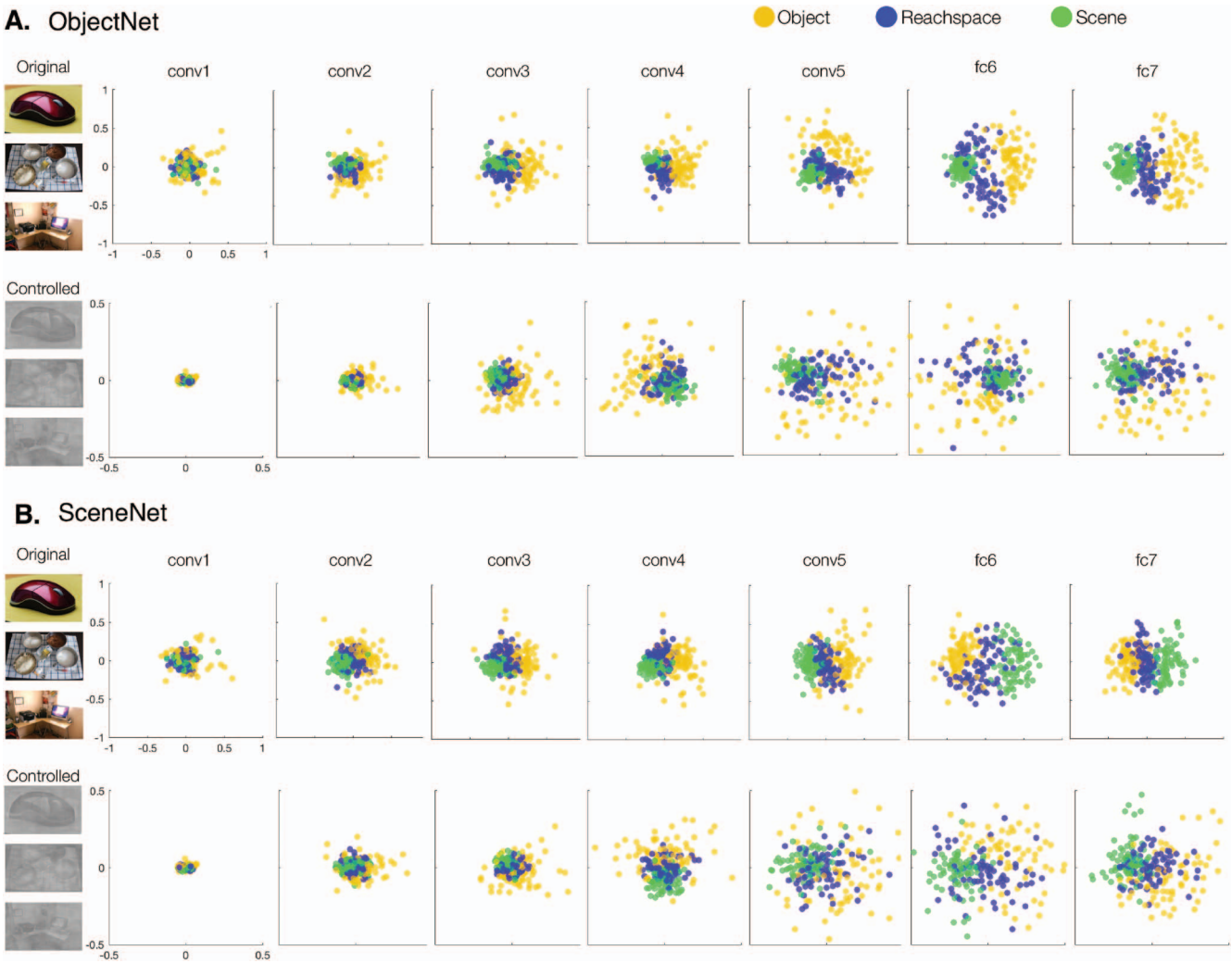

Figure 5. Multidimensional scaling plots of the feature similarity for objects (yellow dots), reachspaces (blue dots), and scenes (green dots). (A) Plots for each of the 7 layers of a deep net trained to do object recognition, where features were extracted from original images (top plot), and controlled images (bottom plot). (B). The same visualizations are shown as in (A) but for a network trained to classify scene categories. Note that the scale is different for the original image subplots compared to the controlled image subplots. The order in which data points were added to the graph was random, and data points are semi-transparent to enable clearer visualization of the distributions for all three scales of space. See appendix for image attributions.

classification approach (Figure 7, top panel). A naïve Bayes classifier was trained to predict whether a held-out image was an object, reachspace, or scene based on the feature activations of each net and each layer. Cross-validated prediction accuracy was above chance in all layers. Note that the naïve Bayes classifier did not show the performance boost for original images in Layer fc6 that is apparent in the data-driven analysis. However, this analysis generally showed a similar pattern of results to the data driven analysis: Images can be classified by scale, and classification accuracy was highest in intermediate and later layers, for both ObjectNet and SceneNet, for both original and controlled image sets.

Finally, we additionally probed a number of other distinctions (see Figure 7, bottom panel). First, we considered whether objects, reachspaces, and scene feature differences would be evident in direct two-way comparisons. For this analysis, we included only two scales of space, and used naïve Bayes to assess how distinguishable pairs of scales are from each other. Overall, we found that objects and scenes are the most easily distinguished from each other but that reachspaces remain highly dissociable from both scenes and objects in two-way comparisons for both original images and controlled images. Second, we examined whether the six sematic categories were distinguishable from each other across in these feature spaces. We found that six-way classification was above chance by later layers, but overall was much less accurate than classification by scale of space. These results are consistent with the finding from Experiment 2 that perceptual differences between reachspaces and scenes were stronger than those for semantic category.

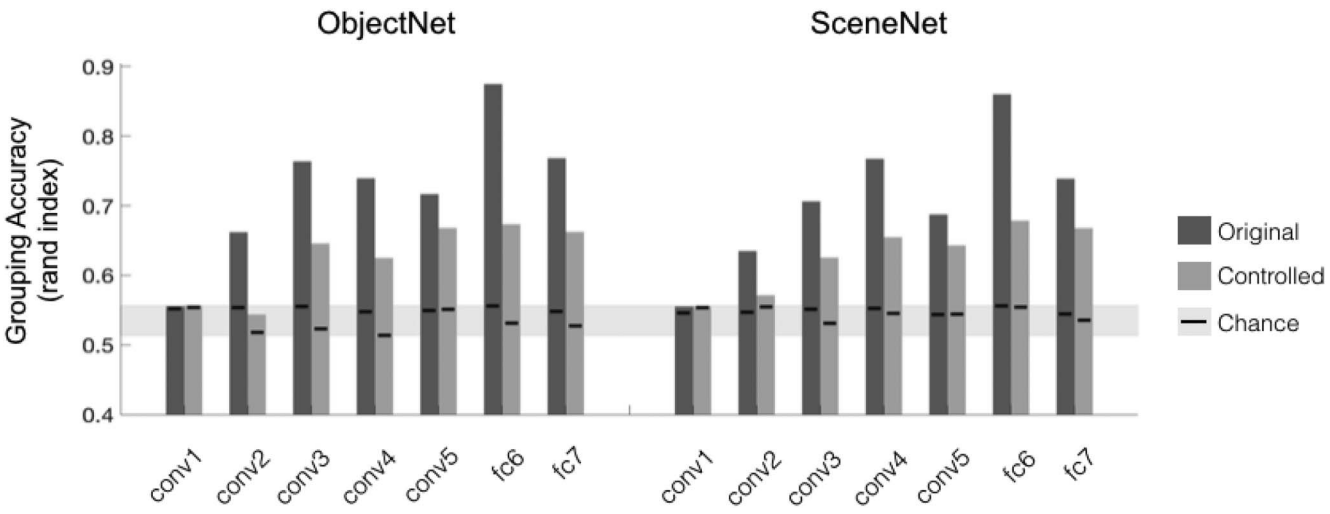

Figure 6. Grouping accuracy results. The grouping accuracy score reflects how well the data-driven clusters recover the object, reachspace, and scene classes, quantified with the Rand index. Grouping accuracy is plotted separately for ObjectNet (left) and SceneNet (right), for each layer, for both original images (dark gray bars) and controlled images (light gray bars). Chance was estimated separately for condition (small black horizontal line on each bar), with the range of chance values across these conditions depicted (light gray bar).

Taken together, these results primarily serve as an existence proof that there are image feature spaces in which reachspaces dissociate from both scenes and objects. However, these results also begin to provide some insight into the nature of the features that differentiate reachspaces from objects and scenes. First, given that grouping accuracy increases over layers, and that the correct grouping is still recovered when differences in luminance and spatial frequency are minimized, it is likely that features of mid- to high-level complexity underlie the divisions among scales of space. Second, the fact that reachspaces dissociated from other scales in both object-trained and scene-trained networks implies that this distinction does not rely solely on features specialized for distinguishing between specific object categories or specific scene categories. Third, the MDS visualization suggest that reachspaces occupy an intermediate position between objects and scenes in these feature spaces, further complementing the patterns in the behavioral data.

General Discussion

Research on the visual representation of the environment has proceeded largely by drawing a division between single objects and full-scale scenes, even though this may be an oversimplification of continuous visual experience (cf. Henderson & Hollingworth, 1999). In the present study, we explicitly move away from this approach and directly ask whether near-scale views of the environment, reachspaces, are perceptually distinct from full-scale scene views as well as singleton object views. Using human behavioral studies and deep neural networks, we found strong evidence for this dissociation. In Experiment 1 of the behavioral studies, visual search patterns showed a three-way dissociation between objects, reachspaces, and scenes for both luminance-matched and spatial-frequency-matched image sets. In Experiment 2, we showed that the perceptual difference between scenes and reachspaces was substantial: The magnitude of perceptual differences between scenes and reachspaces was much larger than the differences between scene categories in this image set. Finally, complementing these patterns in human behavior, we found that the features spaces learned by two different deep neural networks also naturally dissociate reachspace views relative to objects and scene views.

Taken together, the current study demonstrates that there are systematic feature distinctions between reachspaces and scenes, detectable by the human perceptual system, likely related to visual features of intermediate complexity. Although it may be tempting to interpret this dissociation as evidence that reachspaces and scenes are separate mental categories, the present data do not support that inference, and future work will be required to explore this possibility. Rather, the current results point to the existence of systematic differences in *perceptual content* between scales of space previously treated as interchangeable. These results highlight that models of scene perception that do not distinguish between near and far space may be incomplete. In the following sections, we discuss the nature of the feature differences between reachspaces and scenes, the broader construct of reachspaces, and the implications that this perceptual division of space may have for our cognitive and neural architecture.

The Nature of the Visual Feature Distinctions

Given that reachspaces look systematically different from full-scale scenes, what are the visual features supporting this dissociation? We can draw some inferences about these features from the modeling results. Previous work has shown that in deep neural networks trained to do object or scene recognition, feature tuning increases in complexity over successive layers (Güçlü & van Gerven, 2015; Zeiler & Fergus, 2014; Zhou, Lapedriza, Xiao, Torralba, & Oliva, 2014), in which Layer 1 is dominated almost exclusively by representations of simple elements (oriented lines and colors), Layers 2 and 3 see an increase in representations of textures and surfaces alongside the simple elements, and later

Classification Accuracy

comparison	image type	ObjectNet conv1	conv2	conv3	conv4	conv5	fc6	fc7	SceneNet conv1	conv2	conv3	conv4	conv5	fc6	fc7
Three-way classification: how discriminable are the three scales of space?															
RS vs S vs O	raw	60.7	79.6	85.7	84.3	87.0	78.2	85.7	64.8	79.2	82.9	88.9	90.3	81.9	89.8
	controlled	56.9	66.2	80.1	77.3	69.9	71.3	72.2	55.6	67.6	74.5	75.9	70.8	66.7	74.5
Two-way classification: how discriminable are scales of space in pairwise comparisons?															
RS vs S	raw	64.6	86.8	92.4	87.5	90.3	81.9	90.3	69.4	91.0	92.4	94.4	95.8	91.7	95.8
	controlled	61.8	69.4	84.0	77.8	74.3	78.5	78.5	58.3	73.6	77.1	81.9	77.8	76.4	82.6
RS vs O	raw	76.4	84.0	88.2	89.6	91.0	81.9	90.3	77.8	79.9	82.6	89.6	89.6	77.1	88.2
	controlled	72.9	79.2	85.4	88.2	80.6	73.6	82.6	76.4	77.5	84.0	81.9	78.5	71.5	79.9
O vs S	raw	80.6	91.7	94.4	96.5	95.8	93.8	95.8	84.0	93.1	97.2	96.5	96.5	99.3	98.6
	controlled	79.9	84.7	91.0	93.8	89.6	88.2	88.9	79.9	84.0	90.3	91.7	92.4	79.2	95.1
Six-way classification: how discriminable are the six semantic categories, cutting across scales of space?															
Category decoding: O,RS,S included	raw	30.1	44.0	46.8	46.1	48.2	43.5	50.5	31.9	43.5	50.5	52.3	59.8	49.5	55.6
	controlled	26.9	30.6	27.8	25.2	31.9	29.2	34.3	25.5	30.6	31.9	31.9	31.9	28.7	33.8
Category decoding: RS and S only	raw	47.2	59.7	68.1	59.7	58.3	47.9	56.3	48.6	61.1	65.9	72.9	75.7	56.3	64.6
	controlled	32.6	41.7	37.5	43.8	34.7	31.9	33.3	32.6	36.8	45.8	41.0	42.4	35.4	45.1

Legend: chance performance — 100% correct

Figure 7. Auxiliary results, using naïve Bayes classification with leave-one-out cross-validation. The top panel reports the results of three-way classification of the images into the three scales of space (objects, O; reachspaces, RS; and scenes, S), for both raw and controlled images, in a network pretrained to-do object classification or scene classification (chance for this comparison was 33.3%). The middle panel reports two-way classification accuracy, testing the discriminability of pairs of images (chance for these comparisons was 50%). The bottom panel reports six-way classification accuracy for distinguishing among the six semantic categories, when considering all the images together or only the reachspaces and scenes together, for both raw and controlled images, in both networks (chance for these comparisons was 16.6%). Cells in the table are shaded by their accuracy relative to chance for that particular comparison, with chance grouping performance in light gray and 100% accuracy in dark gray.

layers (4 and 5) contain representations of object parts and full objects (Zhou et al., 2014). Given that all layers beyond the first one or two were sensitive to the object-reachspace-scene distinction, this suggests that mid-level features such as texture and surfaces, as well as high-level features such as object parts and entire objects, may largely underlie the dissociation, whereas simple elements such as color, contrast, and oriented lines do not.

There are also fundamental differences in the structure and constituent parts of reachspaces and scenes that may suggest other mid-level perceptual feature differences to explore. For example, reachspaces are dominated by small objects (e.g., bowls, sinks, pots), whereas scenes are dominated by large ones (e.g., tables, desks, rugs). Given that curvature varies with object size (Long et al., 2016; Long, Yu, & Konkle, 2018), it is possible that reachspaces have more curved contours, whereas scenes have more rectilinear contours. Additionally, the reachspaces in this stimulus set all have a bounded surface on the bottom and are open on the top, whereas indoor rooms are enclosed on all sides. Thus, there are likely systematic differences in 3D layout features and the spatial envelope of the views (Oliva & Torralba, 2001, 2006). Finally, full-scene views encompass larger environments, and thus may include more elements and give rise to more perceptual clutter than reachspaces.

One open question is whether reachspaces are encoded in the visual system with distinct perceptual primitives. That is, are reachspaces processed by specific perceptual analyzers that are primarily dedicated to processing near-scale spaces? Or are reachspace processed by weighted combinations of perceptual analyzers devoted to object-specific and scene-specific processing? One potential way to gain insight into this question is to leverage functional neuroimaging: If reachspaces drive some regions along the visual processing stream more strongly than both objects and scenes, this result would favor the possibility

that reachspaces are not simply intermediate and instead may have perceptual features of their own. However, regardless of the nature of the feature distinction, we have shown that for human perception, reachspace views are systematically different than scene views.

Boundary Conditions of Reachspaces

In the present work, we sampled our stimuli from what we think of as "canonical" reachspaces. That is, all reachspace views were of task-relevant near-scale spaces, largely within arm's reach, made up of semantically related objects arrayed on a horizontal surface, drawn from everyday contexts. However, there are other kinds of near-scale spaces with different characteristics. For example, bookshelves, pantries, ATMs, and vending machines are spaces in which we use our hands to manipulate objects, but they are primarily defined by a vertical plane. Likewise, photocopiers, ATMs, and digital kiosks are made up of single large objects rather than multiple discreet objects. Would views of such noncanonical reachspaces still dissociate from scenes and objects, and would they group with more canonical reachspaces? Further, the intermediate perceptual status of our reachspaces relative to objects and scenes also raises an important question: Do reachspaces reflect a point along a continuum from scenes to objects, characterized by smooth changes in perceptual features as the scale of space increases from objects to scenes, or are there more categorical boundaries in which reachspaces reflect a distinct kind, characterized by abrupt changes in perceptual primitives from one scale to the next? Understanding the boundary conditions of what is a reachspace and understanding their categorical status are important new directions that arise from these results.

Implications for Cognitive and Neural Architecture

The evidence for a perceptual dissociation, falling along functionally relevant divisions of space, provides the foundation for future work exploring whether differences in neural and cognitive representations run alongside the perceptual differences reported here. Previous work has shown that patterns of neural activity in occipitotemporal cortex reflect low- and mid-level feature tuning (Groen, Ghebreab, Lamme, & Scholte, 2012; Groen, Silson, & Baker, 2017; Long et al., 2018; Watson, Young, & Andrews, 2016) as well as spatial layout and functional affordances (Bonner & Epstein, 2017; Park, Konkle, & Oliva, 2015). Thus, it is possible that reachspaces will have distinct neural signatures along the ventral stream. Indeed, some previous evidence suggests that scene-processing areas can distinguish between near- and far-scale spaces (Henderson, Larson, & Zhu, 2008; but see Epstein et al., 2003). It is also likely that reachspace views engage cognitive processes that are not engaged by scenes and objects. For example, performing a task in a reachspace requires you to track the state of the task, the properties of the objects that signal the next step, and the possible hand actions to be performed during that step (Hayhoe, 2000; Kirsh, 1995; Triesch, Ballard, Hayhoe, & Sullivan, 2003), while navigating through a scene requires you to track a very different set of attributes. Given that we spend most of our days in environments in which objects are within reach and we are performing tasks, it is surprising that we still know very little about how these environments are represented. This study represents a step in extending our understanding of object and scene perception mechanisms to the perception of reachable space.

References

Bertamini, M., Jones, L. A., Spooner, A., & Hecht, H. (2005). Boundary extension: The role of magnification, object size, context, and binocular information. *Journal of Experimental Psychology: Human Perception and Performance, 31,* 1288–1307. http://dx.doi.org/10.1037/0096-1523.31.6.1288

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94,* 115–147. http://dx.doi.org/10.1037/0033-295X.94.2.115

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

Bonner, M. F., & Epstein, R. A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLoS Computational Biology, 14,* e1006111. http://dx.doi.org/10.1371/journal.pcbi.1006111

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436. http://dx.doi.org/10.1163/156856897X00357

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., . . . DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology, 10,* e1003963. http://dx.doi.org/10.1371/journal.pcbi.1003963

Carey, D. P., Dijkerman, H. C., Murphy, K. J., Goodale, M. A., & Milner, A. D. (2006). Pointing to places and spaces in a patient with visual form agnosia. *Neuropsychologia, 44,* 1584–1594. http://dx.doi.org/10.1016/j.neuropsychologia.2006.01.024

Carey, S., & Xu, F. (2001). Infants' knowledge of objects: Beyond object files and object tracking. *Cognition, 80,* 179–213. http://dx.doi.org/10.1016/S0010-0277(00)00154-2

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

Cohen, M. A., Alvarez, G. A., Nakayama, K., & Konkle, T. (2017). Visual search for object categories is predicted by the representational architecture of high-level visual cortex. *Journal of Neurophysiology, 117,* 388–402. http://dx.doi.org/10.1152/jn.00569.2016

Cowey, A., Small, M., & Ellis, S. (1994). Left visuo-spatial neglect can be worse in far than in near space. *Neuropsychologia, 32,* 1059–1066. http://dx.doi.org/10.1016/0028-3932(94)90152-X

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96,* 433–458. http://dx.doi.org/10.1037/0033-295X.96.3.433

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

Epstein, R., Deyoe, E. A., Press, D. Z., Rosen, A. C., & Kanwisher, N. (2001). Neuropsychological evidence for a topographical learning mechanism in parahippocampal cortex. *Cognitive Neuropsychology, 18,* 481–508. http://dx.doi.org/10.1080/02643290125929

Epstein, R., Graham, K. S., & Downing, P. E. (2003). Viewpoint-specific scene representations in human parahippocampal cortex. *Neuron, 37,* 865–876. http://dx.doi.org/10.1016/S0896-6273(03)00117-X

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

Gallivan, J. P., Cavina-Pratesi, C., & Culham, J. C. (2009). Is that within reach? fMRI reveals that the human superior parieto-occipital cortex encodes objects reachable by the hand. *The Journal of Neuroscience, 29,* 4381–4391. http://dx.doi.org/10.1523/JNEUROSCI.0377-09.2009

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

Greene, M. R., & Oliva, A. (2010). High-level aftereffects to global scene properties. *Journal of Experimental Psychology: Human Perception and Performance, 36,* 1430–1442. http://dx.doi.org/10.1037/a0019058

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research, 41,* 1409–1422. http://dx.doi.org/10.1016/S0042-6989(01)00073-6

Groen, I. I., Ghebreab, S., Lamme, V. A., & Scholte, H. S. (2012). Spatially pooled contrast responses predict neural and perceptual similarity of naturalistic image categories. *PLoS Computational Biology, 8,* e1002726. http://dx.doi.org/10.1371/journal.pcbi.1002726

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife, 7,* e32962. http://dx.doi.org/10.7554/eLife.32962

Groen, I. I., Silson, E. H., & Baker, C. I. (2017). Contributions of low-and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society: Series B, Biological Sciences, 372,* 20160102.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience, 35,* 10005–10014. http://dx.doi.org/10.1523/JNEUROSCI.5023-14.2015

Halligan, P. W., & Marshall, J. C. (1991). Left neglect for near but not far space in man. *Nature, 350,* 498–500. http://dx.doi.org/10.1038/350498a0

Hayhoe, M. (2000). Vision using routines: A functional account of vision. *Visual Cognition, 7,* 43–64. http://dx.doi.org/10.1080/135062800394676

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50,* 243–271. http://dx.doi.org/10.1146/annurev.psych.50.1.243

Henderson, J. M., Larson, C. L., & Zhu, D. C. (2008). Full scenes produce more activation than close-up scenes and scene-diagnostic objects in parahippocampal and retrosplenial cortex: An fMRI study. *Brain and Cognition, 66,* 40–49. http://dx.doi.org/10.1016/j.bandc.2007.05.001

Intraub, H. (2010). Rethinking scene perception: A multisource model. *Psychology of Learning and Motivation, 52,* 231–264. http://dx.doi.org/10.1016/S0079-7421(10)52006-1

Intraub, H. (2012). Rethinking visual scene perception. *Wiley Interdisciplinary Reviews: Cognitive Science, 3,* 117–127.

Intraub, H., Bender, R. S., & Mangels, J. A. (1992). Looking at pictures but remembering scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 180–191. http://dx.doi.org/10.1037/0278-7393.18.1.180

Josephs, E., & Konkle, T. (2018, July 9). *Perceptual dissociations among views of objects, scenes, and reachable spaces*. http://dx.doi.org/10.31234/osf.io/gm43c

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology, 8,* 1726. http://dx.doi.org/10.3389/fpsyg.2017.01726

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology, 10,* e1003915. http://dx.doi.org/10.1371/journal.pcbi.1003915

King, M., Groen, I. I. A., Steel, A., Kravitz, D., & Baker, C. (2018). Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *bioRxiv*. Advance online publication. http://dx.doi.org/10.1101/316554

Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence, 73,* 31–68. http://dx.doi.org/10.1016/0004-3702(94)00017-U

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception, 36,* 1–16.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advance in neural information processing systems* (pp. 1097–1105). Cambridge, MA: MIT Press.

Landis, T., Cummings, J. L., Benson, D. F., & Palmer, E. P. (1986). Loss of topographic familiarity. An environmental agnosia. *Archives of Neurology, 43,* 132–136. http://dx.doi.org/10.1001/archneur.1986.00520020026011

Long, B., Konkle, T., Cohen, M. A., & Alvarez, G. A. (2016). Mid-level perceptual features distinguish objects of different real-world sizes. *Journal of Experimental Psychology: General, 145,* 95–109. http://dx.doi.org/10.1037/xge0000130

Long, B., Störmer, V. S., & Alvarez, G. A. (2017). Mid-level perceptual features contain early cues to animacy. *Journal of Vision, 17*(6), 20. http://dx.doi.org/10.1167/17.6.20

Long, B., Yu, C. P., & Konkle, T. (2018). Mid-level visual features explain the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences, 115,* E9015–E9024. http://dx.doi.org/10.1073/pnas.1719616115

Maravita, A., & Iriki, A. (2004). Tools for the body (schema). *Trends in Cognitive Sciences, 8,* 79–86. http://dx.doi.org/10.1016/j.tics.2003.12.008

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Reason, 4,* 61–64.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42,* 145–175. http://dx.doi.org/10.1023/A:1011139631724

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research, 155,* 23–36. http://dx.doi.org/10.1016/S0079-6123(06)55002-2

Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance, 37,* 58–71. http://dx.doi.org/10.1037/a0020747

Park, S., Konkle, T., & Oliva, A. (2015). Parametric coding of the size and clutter of natural scenes in the human brain. *Cerebral Cortex, 25,* 1792–1805. http://dx.doi.org/10.1093/cercor/bht418

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442. http://dx.doi.org/10.1163/156856897X00366

Potter, M. C. (1975). Meaning in visual search. *Science, 187,* 965–966. http://dx.doi.org/10.1126/science.1145183

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 509–522. http://dx.doi.org/10.1037/0278-7393.2.5.509

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin, 86,* 446–461. http://dx.doi.org/10.1037/0033-2909.86.3.446

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association, 88,* 1273–1283. http://dx.doi.org/10.1080/01621459.1993.10476408

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision, 115,* 211–252. http://dx.doi.org/10.1007/s11263-015-0816-y

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science, 14,* 29–56. http://dx.doi.org/10.1207/s15516709cog1401_3

Steeves, J. K., Humphrey, G. K., Culham, J. C., Menon, R. S., Milner, A. D., & Goodale, M. A. (2004). Behavioral and neuroimaging evidence

for a contribution of color and texture information to scene classification in a patient with visual form agnosia. *Journal of Cognitive Neuroscience, 16,* 955–965. http://dx.doi.org/10.1162/0898929041502715

Torralba, A., & Oliva, A. (2002). Depth estimation from image structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24,* 1226–1238. http://dx.doi.org/10.1109/TPAMI.2002.1033214

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network, 14,* 391–412. http://dx.doi.org/10.1088/0954-898X_14_3_302

Triesch, J., Ballard, D. H., Hayhoe, M. M., & Sullivan, B. T. (2003). What you see is what you need. *Journal of Vision, 3*(1), 9. http://dx.doi.org/10.1167/3.1.9

Võ, M. L. H., & Wolfe, J. M. (2013). Differential electrophysiological signatures of semantic and syntactic scene processing. *Psychological Science, 24,* 1816–1823. http://dx.doi.org/10.1177/0956797613476955

Watson, D. M., Young, A. W., & Andrews, T. J. (2016). Spatial properties of objects predict patterns of neural response in the ventral visual pathway. *NeuroImage, 126,* 173–183. http://dx.doi.org/10.1016/j.neuroimage.2015.11.043

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods, 42,* 671–684. http://dx.doi.org/10.3758/BRM.42.3.671

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour, 1,* 0058. http://dx.doi.org/10.1038/s41562-017-0058

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America, 111,* 8619–8624. http://dx.doi.org/10.1073/pnas.1403112111

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *European conference on computer vision* (pp. 818–833). Cham, Switzerland: Springer.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Neural Information Processing Systems 2014* (pp. 487–495). Cambridge, MA: MIT Press.

Appendix

Sample Size Estimation

All experiments were initially run with 12 subjects each, which, in the past, has been a standard sample size in visual search studies. However, among new efforts to increase reliability and replicability of results, reviewers requested that we add subjects in each experiment to increase power. To estimate a new target sample size, we used a simulation technique to estimate power over a range of sample sizes. We set the desired power value to 80% to detect the difference between scenes-among-scenes and scenes-among-reachspaces (one of our two critical effects).

In this simulation approach, hypothetical data sets for a range of sample sizes were constructed from previously obtained data. Specifically, for a given sample size, subjects were randomly sampled with replacement from the initial 12 subjects in the experiment. Trials were obtained for each of these simulated subjects by randomly sampling trials with replacement from the subject's original trials. The simulated data set had the same number of trials per condition as the source dataset. Next, the average reaction time for each subject in each condition was calculated from this simulated data set, and a paired t test was performed for the critical comparison. This simulation was repeated 1,000 times, and power was calculated as the percent of those simulations where $p < .05$. This process was repeated for each sample size ranging from 10 to 120 participants, and the lowest sample size that achieved at least 80% power was recorded.

Applying this method, we obtained an estimated sample size of 28 from the Experiment 1a data, and 15 subjects from the Experiment 1b data. To derive a final sample size estimate that would be more robust to differences in variability between these two data sets, the samples size estimates were averaged over the two experiments, yielding a final sample size of 22 for both experiments. To estimate sample size for Experiment 2, which has fewer trials per condition than Experiment 1, we repeated the simulation but sampled 30 trials per conditions during the trial-sampling step, matching the design of Experiment 2. Estimates were again derived from Experiments 1a and 1b data separately and were averaged to yield a final sample size of 27 subjects for Experiment 2.

Following these analyses, additional data were collected to bring the total number of participants from 12 to 22 in both Experiments 1a and 1b, and to bring the total participants from 12 to 27 in Experiment 2. The overall patterns in the results were not changed by adding more participants, but all effects were substantially more robust.

Image Attributions

All images used in the paper were public domain images selected to resemble the experimental stimuli. Images were obtained from Flickr.com, PixHere.com, Pexel.com or MaxPixel.net, and all were either under a CC0 or CC2 license (free to use and modify without attribution).

Received July 2, 2018
Revision received November 7, 2018
Accepted December 8, 2018 ■