Rapid ensemble encoding of average scene features

Vignash Tharmaratnam
Graduate Program in Psychology,
University of Toronto, Toronto, ON, Canada

Jason Haberman
Rhodes College, Memphis, TN, USA

Jonathan S. Cant
Graduate Program in Psychology,
University of Toronto, Toronto, ON, Canada
Department of Psychology,
University of Toronto Scarborough,
Toronto, ON, Canada

Visual ensemble perception involves the rapid global extraction of summary statistics (e.g., average features) from groups of items, without requiring single-item recognition and working memory resources. One theory that helps explain global visual perception is the principle of feature diagnosticity. This is when informative bottom-up visual features are preferentially processed to complete the task at hand by being consistent with one's top-down expectations. Past literature has studied ensemble perception using groups of objects and faces and has shown that both low-level (e.g., average color, orientation) and high-level visual statistics (e.g., average crowd animacy, object economic value) can be efficiently extracted. However, no study has explored whether summary statistics can be extracted from stimuli higher in visual complexity, necessitating global, gist-based processing for perception. To investigate this, across five experiments we had participants extract various summary statistical features from ensembles of real-world scenes. We found that average scene content (i.e., perceived naturalness or manufacturedness of scene ensembles) and average spatial boundary (i.e., perceived openness or closedness of scene ensembles) could be rapidly extracted within 125 ms, without reliance on working memory. Interestingly, when we rotated the scenes, average scene orientation could not be extracted, likely because the perception of diagnostic edge information (i.e., cardinal edges for typically encountered upright scenes) was disrupted when rotating the scenes. These results suggest that ensemble perception is a flexible resource that can be used to extract summary statistical information across multiple stimulus types but also has limitations based on the principle of feature diagnosticity in global visual perception.

Introduction

The processing of the abundant visual information available to the brain is often impeded by limitations in one's working memory and selective attention (Dux & Marois, 2010; Luck & Vogel, 2013; Simons & Levin, 1997). To circumvent this, visual processing must take advantage of the inherent redundancy that often exists in our visual field, in the form of repeating spatial features such as global textures and patterns (Field, 1987; Kersten, 1987; Kinchla, 1977). Indeed, when viewing groups of similar visual items (i.e., ensembles), their redundancy allows for a compression of visual information into a statistical summary or average, enabling faster and more efficient processing.

Ensemble perception is the rapid and precise encoding of statistical information (e.g., averages) from multiple items, without the need to recognize any individual item from the set (Whitney & Yamanashi Leib, 2018). Summary statistics from ensembles have been shown to be accurately coded for low-level features such as average color (Maule, Witzel, & Franklin, 2014), size (Ariely, 2001), motion (Watamaniuk & McKee, 1998), and orientation (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001), and also for high-level features such as average facial expression (Haberman & Whitney, 2009) and crowd animacy (Yamanashi Leib, Kosovicheva, & Whitney, 2016). Even though spreading attention to multiple items decreases the resolution and recognition accuracy for each individual item (Ariely, 2001; Palmer, 1990), accuracy remains high for classifications of the whole ensemble, because the noise within each ensemble element is averaged out of the overall percept (Alvarez, 2011). The fidelity of global ensemble summary statistics—at the expense of local

Citation: Tharmaratnam, V., Haberman, J., & Cant, J. S. (2026). Rapid ensemble encoding of average scene features. *Journal of Vision*, *26*(1):3, 1–31, https://doi.org/10.1167/jov.26.1.3.

https://doi.org/10.1167/jov.26.1.3 Received May 24, 2025; published January 5, 2026 ISSN 1534-7362 Copyright 2026 The Authors

precision—supports research suggesting that visual processing is neither strictly bottom-up or top-down, but rather is a specific optimized combination of both that most efficiently encodes the current stimuli at the time (Kinchla & Wolfe, 1979).

The parahippocampal place area (PPA) plays a key role in optimally extracting global ensemble features. For example, in ensembles composed of similar objects, the processing of both the shape and texture features of the ensemble objects occurs within the PPA (Cant & Xu, 2012; Cant & Xu, 2017), and the processing of each is not done independently of each other (Cant, Sun, & Xu, 2015). In addition, the PPA has also been shown to be sensitive to changes in the ratio of two different objects composing an ensemble, where global estimates of the object ensemble's average shape and texture vary (Cant & Xu, 2015). In contrast, the lateral occipital cortex (LOC) has been shown to be sensitive to changes that do not vary the object ensemble's average shape and texture, such as the spatial arrangement of objects within ensembles (with the number and type of ensemble elements remaining fixed; Cant & Xu, 2015), or changes in the density of objects in ensembles (without changing the ratio; Cant & Xu, 2015). Consequently, these studies suggest that the PPA may play a key role in extracting summary statistics for global visual information, while the LOC may be playing a role in coding local visual elements.

In addition to global ensemble features, PPA is heavily involved in scene processing and is sensitive to a wide range of global scene features. For example, PPA is sensitive to textural patterns within a scene (Lowe, Gallivan, Ferber, & Cant, 2016), the layout or geometry of space (Epstein, Graham, & Downing, 2003), a scene's spatial boundary (i.e., openness/closedness; Park, Brady, Greene, & Oliva, 2011), non-scene landmark objects like buildings (Bastin et al., 2013; Cate, Goodale, & Köhler, 2011), the object content of a scene (Harel, Kravitz, & Baker, 2013), the category a scene belongs to (artificial/natural; Walther, Caddigan, Fei-Fei, & Beck, 2009), the contour junction statistics of scenes (Choo & Walther, 2016), and the spatial frequency content of scenes (Berman, Golomb, & Walther, 2017), as well as apparent scene temperature (i.e., how hot or cold a scene looks) and sound level (i.e., how quiet or noisy a scene appears; Jung & Walther, 2021). Similar to ensemble perception, global scene features are extracted rapidly (as fast as 100 ms; Oliva, 2005; Potter & Faulconer, 1975), circumventing limitations of attention and working memory (McNair, Goodbourn, Shone, & Harris, 2017; Oliva, 2005). These findings as a whole indicate that PPA is incredibly versatile in rapidly extracting and compressing global statistical visual information, as evidenced by the shared neural mechanisms across ensemble and scene perception in parahippocampal cortex (Cant & Xu, 2012; Cant & Xu, 2015; Cant & Xu, 2017; Cant & Xu, 2020).

Traditionally, the neural mechanisms underlying global visual scene processing (e.g., in the PPA) was assumed to follow the "coarse-to-fine hypothesis" (Navon, 1977). This theory suggests that low spatial frequency (LSF) scene content (i.e., the coarser gradients and textural patterns in scenes) is processed first to provide a rough estimate of scene "gist", which is then followed by the processing of high spatial frequency (HSF) content (i.e., the detailed borders, contours and edges in scenes) to provide finer details (Schyns & Oliva, 1997). Although elegant in its simplicity, the coarse-to-fine hypothesis received mixed support in the neuroimaging literature, particularly in PPA. For example, some studies suggested PPA activates more strongly to LSF than HSF content (Peyrin, Baciu, Segebarth, & Marendaz, 2004), while others demonstrated that PPA preferentially processes HSF information (Rajimehr, Devaney, Bilenko, Young, & Tootell, 2011; Zeidman, Mullally, Schwarzkopf, & Maguire, 2012).

A modern theory that can help explain these discrepant results, as well as the versatility of scene-processing mechanisms, is the principle of feature diagnosticity. This theory argues that visual features that are most informative to perform a given task are preferentially used by the visual system, given the available visual information. This is achieved through an interplay of bottom-up sensory inputs and top-down expectations shaped by prior knowledge (Greene & Oliva, 2009a; Oliva & Schyns, 1997). Specifically, bottom-up processing in early visual areas (V1/V2) detects low-level cues like spatial frequencies and edge statistics (Hubel & Wiesel, 1968), while top-down processing in scene-selective regions (e.g., PPA, the occipital place area (OPA), the retrosplenial complex (RSC)) integrates these low-level cues with expectations that bias feature selection (Bar, 2004; Dilks, Julian, Paunov, & Kanwisher, 2013; Miller, Vedder, Law, & Smith, 2014; Park et al., 2011). For example, when evaluating scene content (i.e., how natural or manufactured a scene appears), informative bottom-up cues for natural scenes (e.g., forests, beaches, etc.) include LSFs and complex, irregular textures, whereas informative low-level cues for manufactured scenes (e.g., offices, cities) include HSFs and cardinal edge orientations (Geisler, 2008; Greene & Oliva, 2009a; Walther & Shen, 2014). Top-down expectations (shaped by lifelong statistical learning mechanisms that track correlations between particular visual cues and certain environmental settings; Geisler, 2008) that anticipate more textured natural scenes and more structured manufactured settings, contribute to scene processing by prioritizing these low-level visual cues when evaluating scene naturalness (Bar, 2004; Greene & Oliva, 2009a). These features become diagnostic to completing the task at hand, allowing for rapid global visual processing. Indeed, neuroimaging findings directly support the theory of feature diagnosticity for scene processing. For example, Lowe and colleagues (2016) found that during scene classification, the PPA is more sensitive to variations in scene layout than texture in manufactured scenes, but is equally sensitive to variations in spatial layout and texture in natural scenes.

When evaluating another global scene property, spatial boundary (i.e., how open or closed a scene appears), informative bottom-up cues involve smooth spatial frequency gradients and sparse edge content in open scenes (e.g., plains), and abrupt high-frequency transitions and dense edge structures (e.g., T-junctions) in closed scenes (e.g., caves) (Park et al., 2011; Walther & Shen, 2014). According to the principle of feature diagnosticity, our top-down expectations of vast horizons for open scenes and confined spaces for closed scenes, processed in PPA, OPA, and RSC, serve to prioritize these low-level cues when evaluating scene openness (Dilks et al., 2013; Miller et al., 2014; Park et al., 2011). By using top-down expectations to prioritize the processing of certain diagnostic features, the visual system is thus able to manage the computational load associated with scene processing (Oliva, 2005; Oliva & Schyns, 1997).

Although the principle of feature diagnosticity has been explored for its role in the recognition of single scenes (Harel et al., 2020; Lowe et al., 2016), its role in processing multiple scenes (i.e., scene ensembles) remains unexamined. Scene ensemble perception has strong ecological implications, as the brain integrates multiple scenes both temporally and spatially in real-world contexts (Epstein & Baker, 2019). The human brain has been shown to robustly anticipate scene information as we navigate through different environments (Shikauchi & Ishii, 2016). During navigation, different scenes are encountered sequentially, such as transitioning from an open natural city park to a closed manufactured urban landscape. As we move through different environments, the visual system encodes average changes in diagnostic low-level features comprising scene content and spatial boundaries (Epstein & Baker, 2019; Shikauchi & Ishii, 2016). The average scene features extracted provide rapid yet accurate information to be integrated in real time, even at the expense of detailed information about single isolated elements (Alvarez, 2011; Haberman & Whitney, 2009; Whitney & Yamanashi Leib, 2018). This allows for the generation of more precise expectations of upcoming environments, which can in turn can facilitate later scene gist extraction (Albrecht & Scholl, 2010; Epstein & Baker, 2019; McLean, Nuthmann, Renoult, & Malcolm, 2023). We also encounter scenarios in daily life where multiple scenes are processed simultaneously, such as in contexts when viewing multiple camera feeds on several security monitors or browsing thumbnail images in online search results (e.g., evaluating natural vs. urban scenes on Google Images). In these scenarios, the extraction of average diagnostic features aids rapid judgments about their relevance or context (Greene & Oliva, 2009a; Tiurina, Markov, Whitney, & Pascucci, 2024). Interestingly, both temporal (i.e., multiple sequentially presented items) and spatial (i.e., multiple simultaneously presented items) ensemble processing rely on similar mechanisms, because the brain efficiently averages visual features across time or space with comparable precision (Haberman & Whitney, 2009; Haberman & Whitney, 2011; Khayat, Pavlovskaya, & Hochstein, 2024). By preferentially processing (e.g., attending to) the most diagnostic information relevant for navigation and rapid decision-making processes, the visual system is able to minimize cognitive load, while achieving robust, real-time analysis in complex, dynamic environments (Barhorst-Cates, Rand, & Creem-Regehr, 2020; Epstein & Baker, 2019; Greene & Oliva, 2009a).

Furthermore, individual scene processing is likely subserved at least partially by global ensemble statistical mechanisms, lending support to the idea that ensembles of multiple scenes may also be efficiently processed by these mechanisms. Indeed, Brady, Shafer-Skelton, and Alvarez (2017) showed that one's ability to recognize individual scenes was positively correlated with one's ability to detect global changes in spatial ensembles composed of Gabor elements (i.e., a global ensemble texture), but not with the ability to detect changes in object-based summary statistics. These findings suggest that the processing of global ensemble textures likely contributes to individual scene perception. This is consistent with findings that cortical regions that are sensitive to processing scenes have also been implicated in the processing of textural features of ensembles (Cant & Xu, 2017). Given this, it stands to reason that the global ensemble statistical mechanisms employed during single scene perception could extend to ensembles of multiple scenes. If this is the case, we posit that the principle of feature diagnosticity would underlie this ability.

Indeed, the principle of feature diagnosticity provides a framework for the rapid encoding of statistical information across multiple scenes, despite their abundant visual information. However, the visual complexity of scene ensembles may still pose unique challenges to ensemble processing not seen with other stimuli. When using visual stimuli like objects or faces, ensemble perception remains stable with larger set sizes (see Alvarez, 2011) due to efficient averaging of low-level features (e.g., size (Chong & Treisman, 2003), orientation (Parkes et al., 2001), and facial expression (Haberman & Whitney, 2009)). However, averaging scene information requires the integration of high-level features, such as scene content, spatial boundaries, and scene category information, which impose greater computational demands on the visual system (Cichy et al., 2016; Greene & Oliva, 2009a; Oliva & Torralba, 2006). Consistent with this, Park, Chun, and Johnson (2010) found that the PPA exhibits working memory capacity limitations, with significantly reduced BOLD activation and decoding accuracy when maintaining two scene views compared to one under high cognitive load, indicating a bottleneck in processing multiple scenes simultaneously. Moreover, the OPA primarily processes dynamic, local visual information relevant for moment–to–moment navigation within individual scenes and does not integrate broader spatial context across multiple scenes (Kamps, Julian, Kubilius, Kanwisher, & Dilks, 2016). In contrast, the RSC supports higher–order spatial memory and map–based navigation requiring sustained exposure or temporal integration, constraining its ability to rapidly integrate spatial context across multiple scenes when presented briefly (Alexander, Place, Starrett, Chrastil, & Nitz, 2023). With all this in mind, when encoding scene ensembles, the increased cognitive load could hinder task performance as set size increases, as attentional and neural resources have difficulty extracting multiple features across multiple scenes (Cant & Xu, 2012; Greene & Oliva, 2009a). In contrast, by prioritizing diagnostic features when processing multiple scenes, ensemble-processing mechanisms may be flexible enough to circumvent these aforementioned cognitive limitations and thereby maintain task performance as set-size increases, even across complex scene stimuli.

Moving beyond low-level features of objects, ensemble processing has been shown to be highly flexible, as high-level features such as average crowd animacy and average economic value can be extracted from groups of objects (Yamanashi Leib et al., 2016;Yamanashi Leib, Chang, Xia, Peng, & Whitney, 2020). This establishes that summary statistics can be formed from abstract object features, but it is unclear if this ability extends to the integration of multiple features from multidimensional environments (i.e., containing multiple objects in spatial relation to each other within a particular setting). With this in mind, the goal of the current study was to explore potential limits of ensemble perception by using stimuli that are more visually complex than those used in previous studies. That is, we used ensembles of real-world scenes instead of objects and faces, and investigated whether the principle of feature diagnosticity would mediate this type of ensemble processing, by prioritizing diagnostic low-level features consistent with top-down expectations. Across five experiments, we asked participants to extract estimates of average scene content (i.e., perceived naturalness/manufacturedness), spatial boundary (i.e., perceived openness/closedness), and orientation (i.e., rotated scenes) from scene ensembles. As stated previously, diagnostic features for extracting naturalness and openness include LSF information and complex, irregular textures for natural scenes (e.g., forests, beaches) and smooth spatial frequency gradients and sparse edge content for open scenes (e.g., plains) (Bar, 2004; Brady & Shafer-Skelton, 2017; Greene & Oliva, 2009a; Oliva, Park, & Konkle, 2011; Walther & Shen, 2014). Given the presence of these types of diagnostic visual features in our scene stimuli, we predicted that observers would be able to rapidly, accurately, and globally extract average scene content and spatial boundary from scene ensembles due to the principle of features diagnosticity (i.e., because they align with top-down expectations of textured natural environments and expansive layouts).

There are also visual features that are diagnostic to scene orientation processing, as orientation-selective neurons in early visual cortex extract edge orientations that provide critical information for interpreting spatial layout and scene structure (Hubel & Wiesel, 1968). In typically encountered upright scenes, this comes in the form of cardinal (i.e., horizontal and vertical) edges such as horizons and building outlines. However, in the present study we used rotated scenes, where oblique edges are instead more informative to determine scene orientation (Girshick, Landy, & Simoncelli, 2011; Hubel & Wiesel, 1968; Nasr & Tootell, 2012; Oliva et al., 2011; Shapley & Tolhurst, 1973). This goes against the top-down expectation of using cardinal edges to extract scene orientation. By rotating the scenes, we created a mismatch between the informative oblique edge orientation cues in our scene ensembles and the top-down expectations of prioritizing cardinal edges (Bar, 2004; Charlton, Młynarski, Bai, Hermundstad, & Goris, 2023; Girshick et al., 2011; Greene & Oliva, 2009a). Because of this mismatch, the oblique edges could not be used as diagnostic cues to complete the task, and so we predicted participants would have difficulty extracting average orientation from our scene ensembles.

Finally, we also tested whether these ensemble processing abilities can be explained by appealing to working memory resources. Previous findings demonstrate that ensemble perception does not require single item identification and can occur at a temporal resolution at or beyond the temporal resolution of individual object recognition (Corbett & Oriet, 2011; Haberman & Whitney, 2009), This suggests that ensemble perception operates efficiently by bypassing the limitations inherent to working memory. While some prior studies have considered potential interactions between working memory and ensemble perception (Knox, Pratt, & Cant, 2024; Williams, Pratt, Ferber, & Cant, 2021), none have directly examined interactions between these mechanisms when using complex scene ensembles. Based on these previous results, we predict that average scene content and spatial boundary will be extracted without reliance on working memory resources.

General methods

Participants

Fifty six participants (eight males; 48 females; 51 right-handed; five left-handed; mean age = 18.91 years; age range, 18–22 years) participated in Experiment 1 (content and spatial boundary ratings of individual scenes), 55 participants (18 males; 37 females; 50 right-handed; five left-handed; mean age = 18.85 years; age range, 18–27 years) participated in Experiment 2 (scene ensemble ratings while varying stimulus duration), 42 participants (18 males; 24 females; 38 right-handed; four left-handed; mean age = 20.69 years; age range, 18–34 years) participated in Experiment 3 (scene ensemble ratings while measuring working memory capacity), 49 participants (15 males; 34 females; 45 right-handed; four left-handed; mean age = 19.65 years; age range, 17–29 years) participated in Experiment 4 (orientation ratings of individual scenes), and 61 participants (22 males; 39 females; 56 right-handed; five left-handed; mean age = 19.31 years; age range = 18–27 years) participated in Experiment 5 (scene ensemble orientation ratings). For participant exclusions, see "Data Analysis" below. The participants were all chosen from a pool of undergraduate students at the University of Toronto Scarborough. Participants had normal or corrected-to-normal vision and received course credit for participation. All participants provided informed consent, and the study was approved by the University of Toronto Research Ethics Review Board. Sample sizes (i.e., $n = 42$–61 across experiments) were based on prior ensemble perception studies (e.g., Haberman & Whitney, 2009, with typical samples of 10–20 participants) and a formal power analysis conducted using G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007). For our mixed model analysis (approximated as a repeated measures ANOVA for power estimation), assuming a medium effect size ($f = 0.20$), statistical power of 0.8, and an alpha level of 0.05, power calculations indicated a minimum required sample size of approximately 36 participants. Because the sample size in each of our five experiments exceeds this minimum, we are confident that we have sufficient statistical sensitivity to detect the expected effects in our study.

Stimuli and apparatus

We obtained stimuli from the Places Dataset (http://places.csail.mit.edu/), the SUN database (https://groups.csail.mit.edu/vision/SUN/), and also used stimuli from Oliva and colleagues (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010; Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2017). The stimuli consisted of a wide array of scenes that varied greatly in both scene content (how natural or manufactured a scene appeared) and spatial boundary (how open or closed a scene appeared), and were presented against a white background. The types of stimuli included—but were not limited to—beaches, highways, city buildings, forests, and caves (see Figure 1). All stimuli were size-adjusted using Adobe Photoshop CC 2015 (Adobe Systems Incorporated, San Jose, CA, USA) and were presented using the Psychophysics Toolbox 3 (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997) in MATLAB. Data analysis was conducted using R software (R Core Team, 2022) and MATLAB. Participants were tested in a darkened room and had their head position secured using a headrest that was 32 cm from the front of the table and 40 cm from a CRT monitor (1920 × 1080 pixels, screen refresh rate: 60 Hz). Responses for scene content and spatial boundary ratings were collected through keyboard input, and responses for orientation were collected by mouse input (see Procedure).

In Experiments 1–3 (rating scene content and spatial boundary), each individual scene was sized as a 256 × 256-pixel square. In Experiments 4 and 5 (rating orientation), the individual scenes used in Experiments 1–3 were cut out of the square images by a 256-pixel circle superimposed and centered at the center of the square, to form 256-pixel diameter circular scenes (as if looking at the square individual scenes through a pinhole, see Figure 1). These scenes were cut into circles so that the rectangular frame edges could not act as external orientation cues during the scene orientation rating task (Anderson, Bischof, Foulsham, & Kingstone, 2020; May & Zhaoping, 2016). In Experiments 1 and 4, the individual scenes were positioned at the center of the screen. In Experiments 2, 3 and 5, the set of individual whole scenes to form an ensemble or subset of an ensemble was arranged on a 768 × 512-pixel grid, with the stimuli arranged in three columns by two rows (horizontally subtends 67.7°, vertically subtends 43.3°; see Figures 2 and 3). The location of each stimulus within this grid was randomly selected on each trial.

In Experiment 1, 313 individual scenes were rated for both their scene content and spatial boundary (See Data Analysis; mean, minimum and maximum luminance across pixels: 113.8, 30.7, and 184.6 respectively; values are unitless intensity levels from a uint8 grayscale image, where the pixel luminance ranges from 0 [black] to 255 [white]). After the individual ratings in Experiment 1, four scene ensembles with a set size of six were generated for Experiments 2 and 3 (i.e., whole scene ensembles), wherein participants provided separate ratings of average scene content and spatial boundary for each ensemble, for all possible combinations of one-, two-, four-, and six-item presentations (see Data Analysis). In Experiment 4, the same 313 images (plus

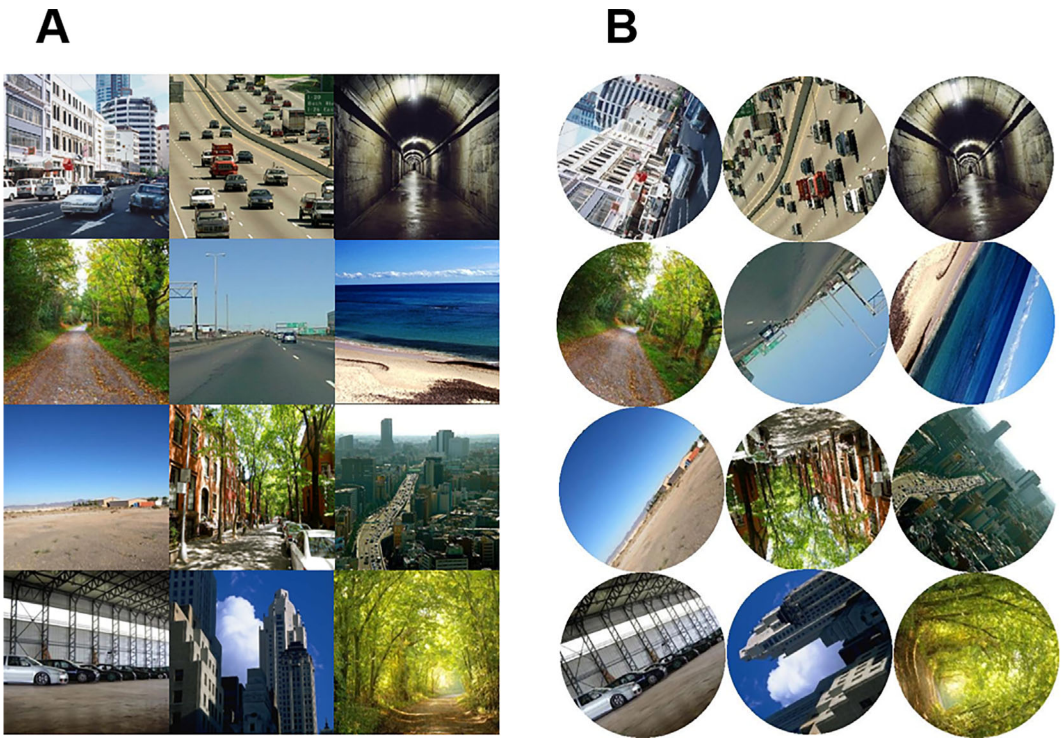

Figure 1. Stimuli used in the study. (**A**) displays examples of scene stimuli presented in Experiments 1–3, illustrating a wide range of variations in scene content (manufactured vs. natural) and spatial boundary (open vs. closed). (**B**) shows the same stimuli converted into circular scenes at various orientations for Experiments 4 and 5.

215 additional scene stimuli) were modified into circular scenes and were then rotated at random degrees so that participants could give orientation responses (see Data Analysis; mean, minimum and maximum luminance across pixels: 119.1, 21.7, and 195.6 respectively). After selecting the stimuli in Experiment 4 that had the most reliable ratings of orientation (see Data Analysis), eight scene ensembles of set size six were generated for Experiment 5, in which participants provided ratings of average orientation.

Procedure

Experiment 1

The procedure for Experiment 1 was adapted from Yamanashi Leib et al. (2016) (also see Alwis and Haberman (2020)). Participants were asked to provide ratings of scene content and spatial boundary for 313 individual scenes in two separate blocks. Ratings were performed on an 11-point integer-based Likert scale, ranging from −5 to 5. The order of the type of rating made, as well as the magnitude of features along the Likert scale (i.e., −5 representing highly manufactured scenes and 5 representing highly natural scenes for scene content ratings, and −5 being highly closed and 5 being highly open for spatial boundary ratings) were counterbalanced between blocks and between participants. The presentation of stimuli within a block was randomized until all 313 images were rated for a given scene feature (313 × 2 = 626 total ratings). Stimuli were presented for one second and were then followed by the presentation of the Likert scale on the screen prompting a rating for a scene feature (see Figure 2). Participants had up to 10 seconds to respond, before the experiment moved on to the next trial. The stimuli were separated by an interstimulus interval (ISI), indicated by a fixation point at the center of the screen that lasted for 500 ms. This trial structure was repeated until all stimuli within a block were presented, and participants were given up to a one-minute break every 40 trials.

Experiment 2

In Experiment 2, four scene ensembles with a set size of six were generated and made up of six pseudorandomly drawn individual scenes (obtained from Experiment 1 without repetition) to design ensembles that required global visual processing to obtain the ensemble average (see Data Analysis below). In addition to making either average scene content or spatial boundary ratings on the whole 6-scene ensemble, participants also rated subsets of either one, two, or four scenes randomly extracted from the

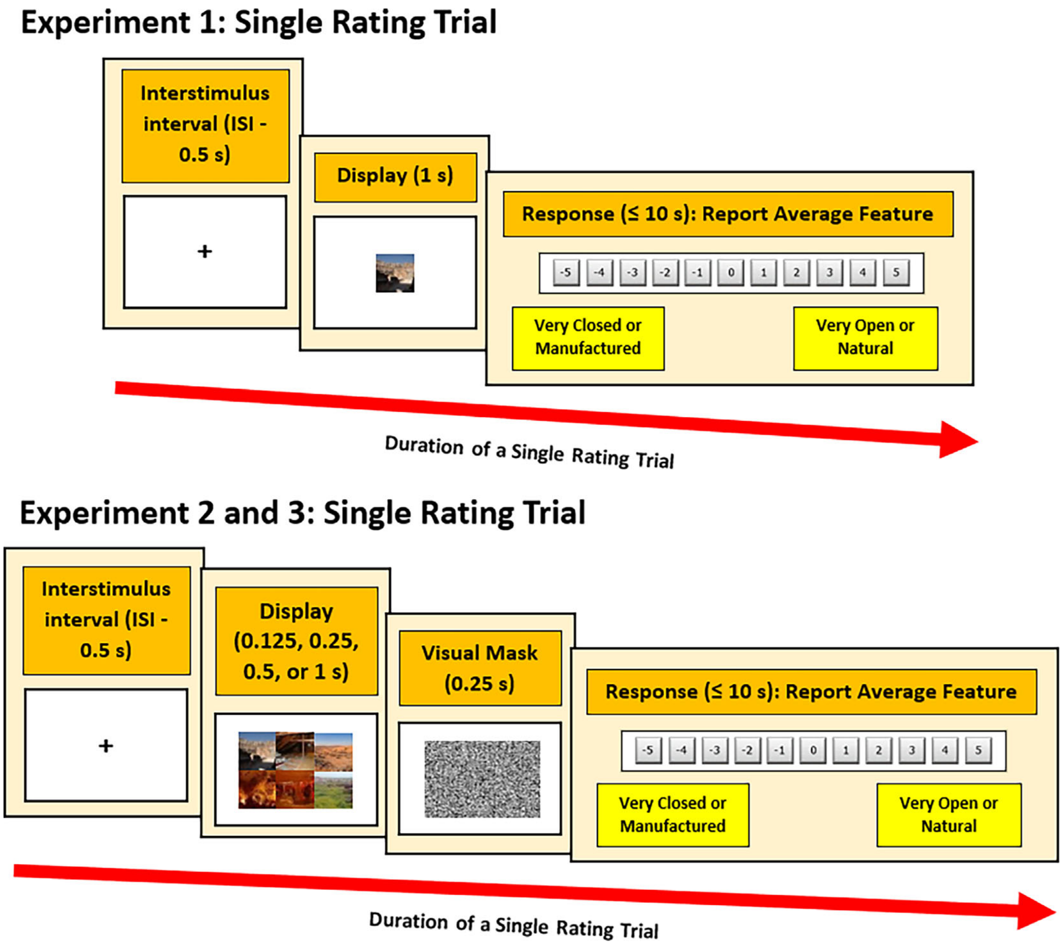
Experiment 1: Single Rating Trial
Interstimulus interval (ISI - 0.5 s)
+
Display (1 s)
Response (≤ 10 s): Report Average Feature
-5 -4 -3 -2 -1 0 1 2 3 4 5
Very Closed or Manufactured
Very Open or Natural
Duration of a Single Rating Trial
Experiment 2 and 3: Single Rating Trial
Interstimulus interval (ISI - 0.5 s)
+
Display (0.125, 0.25, 0.5, or 1 s)
Visual Mask (0.25 s)
Response (≤ 10 s): Report Average Feature
-5 -4 -3 -2 -1 0 1 2 3 4 5
Very Closed or Manufactured
Very Open or Natural
Duration of a Single Rating Trial

Figure 2. Experimental design for a single rating trial in Experiments 1, 2, and 3. At the beginning of every trial, an ISI fixation cross was presented for 0.5 second. In Experiment 1, the ISI was followed by a single scene image presented in the center of the screen for 1 second. In Experiment 2 and Experiment 3's rating trials, the ISI was followed by a one-, two-, four-, or six-item scene ensemble arranged within a 512 × 768 pixel grid. In Experiment 2, the stimulus presentation was for variable amounts of time (0.125 second, 0.250 second, 0.500 second, or 1 second), whereas the stimulus presentation time was 0.250 second in Experiment 3. For Experiments 2 and 3, all stimulus presentations were followed by a subsequent backward mask for 0.25 second. After the stimulus presentation (and masking where applicable), participants were asked to make a response within 10 seconds. For rating trial responses in Experiments 1, 2, and 3, participants were asked to rate either the single (Experiment 1) or average scene content (i.e., natural vs. manufactured) or spatial boundary (i.e., open vs. closed) (Experiments 2 and 3) of the scene(s) presented on a Likert scale ranging from −5 to 5.

whole set scene ensembles. The locations of the items presented were always random within a 3-column × 2-row item grid located at the center of the screen (see Figure 2). All possible combinations for each subset were displayed, leading to six, 15, 15, and one total combination(s) for set sizes of one, two, four, and six items, respectively. To balance the number of trials for each subset, one-item subsets were repeated twice and six-item subsets were repeated 12 times, leading to 12, 15, 15, and 12 trials for the one-, two-, four-, and six-item subsets, respectively. All possible combinations of one-, two-, four-, and six-item presentations for each ensemble were shown to participants, with repetitions, to maintain a similar number of datapoints across set-sizes and to reduce between-subjects noise in the data. Any potential effects of repetition were accounted for in the mixed model used for analysis (see Data Analysis). The stimulus duration was varied between 125, 250, 500, or 1000 ms between blocks, and the presentation of stimuli were followed by a backward mask that lasted 250 ms (see Figure 2). This led to a total of 2 (scene feature rated) × 4 (number of ensembles) × 54 (total number of set-size combinations presented) x 4 (stimulus duration) = 1728 trials in an experimental session. Each feature rating by stimulus duration combination was presented in separate blocks throughout the experiment (4 × 2 = 8 blocks in total, with 216 trials in each block and up to a one-minute break every 40 trials), with the whole set and subset conditions, as well as the ensembles presented within

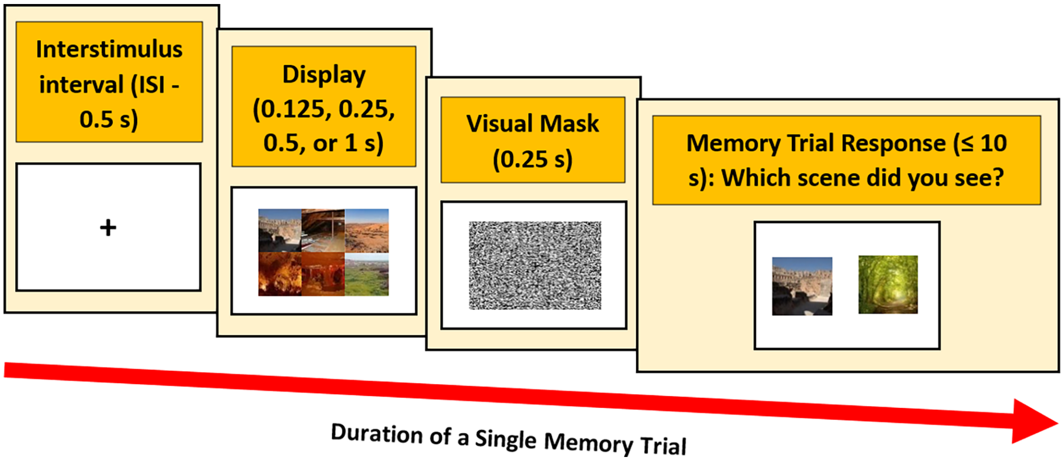
Interstimulus interval (ISI - 0.5 s)
+
Display (0.125, 0.25, 0.5, or 1 s)
Visual Mask (0.25 s)
Memory Trial Response (≤ 10 s): Which scene did you see?
Duration of a Single Memory Trial

Figure 3. Experimental design for a single memory trial in Experiment 3. At the beginning of every trial, an ISI fixation cross was presented for 0.5 second. The fixation cross was followed by six 256 × 256 pixel scene images arranged within a 512 × 768 pixel grid for 0.250 second. After the stimulus presentation and a subsequent backward mask for 0.250 second, one of the six images was displayed again along with an image that was not previously presented. Participants had up to 10 seconds to indicate which of the two images they had seen.

each rating condition, being randomly presented within each block. The order of the blocks was randomized across participants. The temporal structure of each trial is otherwise the same as that described in Experiment 1.

Experiment 3

The same 4 scene ensembles used in Experiment 2 were again in Experiment 3, and the experiments were identical except for two changes. The first change was that there was only one stimulus duration of 250 ms for scene feature rating trials. The second change was, in addition to having scene feature rating trials, an additional block of memory trials was added. In the memory trials, participants performed a two-alternative forced choice task to measure their working memory capacity for the items presented in the ensembles (see Figure 3). At the beginning of every trial, an ISI of 0.5 second was followed by a six-item whole-scene ensemble for 250 ms (no subsets were shown for the memory trials). After the stimulus presentation and a subsequent backward mask for 250 ms, one of the six images presented was displayed (i.e., the target) along with an image that was not presented (i.e., the lure). Participants had up to 10 seconds to indicate (via the left or right arrow key on the keyboard) which of the two images they had seen previously, and had up to a 1-minute break every 40 trials. Each single stimulus within all four scene ensembles was selected across trials four times, leading to 4 × 6 × 4 = 96 memory trials.

Experiments 4 and 5

The procedures for Experiments 4 and 5 are identical to the procedures of Experiments 1 and 3, respectively, except for a few main changes. First, instead of using square scenes, circular scenes were used (see Figure 1). Secondly, in Experiment 4, 215 additional scene stimuli were added to the existing 313 stimuli used in Experiment 1, for a total of 528 stimuli. Third, in report trials participants reported the average orientation of the scene(s) on the screen, by moving the mouse to rotate an arrow on the screen and then confirming the correct orientation with a mouse click.

Data analysis

Determining individual scene content and spatial boundary values

For Experiment 1 (rating individual scenes), in order determine the average scene content or spatial boundary rating for each item across participants, ratings that were more than 1.5 times the interquartile range (IQR) above or below the third and first quartile rating value were excluded. After adjusting for counterbalancing done during the experiment, spatial boundary ratings had −5 representing very closed scenes, 5 representing very open scenes, and 0 representing equally open and closed scenes; while scene content ratings had −5 representing highly manufactured scenes, 5 representing highly natural scenes, and 0 representing equally manufactured and natural scenes. We then wanted to organize individual scene stimuli based on both their scene content and spatial boundary ratings. For each scene feature, 5 bin value boundaries were created: −5 to −3, −3 to −1, −1 to 1, 1 to 3, and 3 to 5. Across both features (scene content and spatial boundary), these boundaries led to 5 × 5 = 25 unique bins that individual scene content and spatial boundary values could occupy. These bins were created to later assemble scene ensembles that were uncorrelated in their average scene content and spatial boundary values. To confirm interrater reliability, a two-way mixed intraclass correlation coefficient was calculated for absolute agreement between every possible pair of participants, for both the scene content and the spatial boundary stimulus sets (McGraw & Wong, 1996).

Composing scene content and spatial boundary scene ensembles

To compose the scene ensembles for Experiments 2 and 3, images were drawn pseudorandomly to form ensembles with skewed distributions of individual scene content and spatial boundary values. The four ensembles were targeted to have a mean scene content and spatial boundary value of 3.5 and 3.5, −3.5 and 3.5, 3.5 and −3.5, and −3.5 and −3.5, respectively. These mean scene feature values for each ensemble were chosen so that across ensembles, mean scene content and spatial boundary values would be uncorrelated with each other. Skewed distributions of scene feature values within an ensemble were used (instead of using a uniform or normal distribution), so that it would be harder to distinguish the ensemble's average scene feature value by allocating attention on only a subset of the items. For instance, imagine a scenario where a participant is presented with a scene ensemble containing 6 images that were randomly selected from a uniform distribution, with scene naturalness values of −5, −3, −1, 1, 3, and 5, yielding an average naturalness value of 0. If the participant used a subsampling strategy when presented with all six scenes, they might pay attention to two of the scenes with a value of −5 and 5, for example, and arrive at an average of 0 (i.e., the same value as the average for all six scenes). This highlights an issue with using uniform or normal distributions in ensemble tasks. That is, due to the symmetry of these distributions, the averages of randomly drawn smaller samples are prone to regression to the mean, which allows for responses made using a subsampling strategy to be mathematically similar in effectiveness compared with a global integration strategy (i.e. generating an average using all available items). Consequently, this would render the results difficult to interpret with regard to ensemble integration. In contrast, using a skewed distribution (e.g., values of −5, 0, 0, 3, 4, and 5) guards against this possibility by making it less likely that the average of a subset approximates the average of the entire set, and thereby encourages the global integration of all items to complete the task.

To make the skewed distributions, for each ensemble, we selected one third of the scenes from bins that most closely matched the target scene feature values, and then randomly drew scenes from all bins for the other two thirds. To verify that this led to individual scene values with high skew (promoting global attention to the ensemble), a simulation of ideal observers was conducted for scene content and spatial boundary ratings (see Supplemental Materials and Figures). Pictures with the lowest standard deviation in their scene feature ratings were selected from bins first, and were not selected more than once. For Experiment 3, working memory capacity was calculated using the following equation: $Working\ Memory\ Capacity = (\%_{Correct} - 0.5)*2*6$ (Yamanashi Leib et al., 2016).

Validating scenes that had reliable orientations

Experiment 4 was performed in the same manner as Experiment 1 except circular scenes were presented instead of square scenes. The purpose of Experiment 4 was to assess how difficult it was for participants to report scene orientation and from this to derive a stimulus set of scenes that could be used in Experiment 5. Specifically, we wanted to examine how reliably participants could perceive the orientation of individual scenes that were rotated through 360° (e.g., if a landscape was made up of only sand dunes without any horizon, orientation information would not be easily discernable and thus participants would likely not reliably perceive its orientation). The scenes were placed behind a circular aperture so that the bounding edges of the original rectangular stimulus could not be used to ascertain the correct orientation. In Experiment 4 each of the individual circular scenes was presented at a random orientation through 360° on each trial. Next, an error measurement (which will be called "orientation difficulty") was calculated for each stimulus by subtracting the absolute difference between the participants' reported orientation and the correct orientation. The higher the orientation difficulty for a stimulus, the harder it is for participants to rate its orientation. These values were averaged for each stimulus across participants to obtain an overall orientation difficulty value for each stimulus. To confirm interrater reliability (similar to Experiment 1), a two-way mixed intraclass correlation coefficient was calculated for absolute agreement between every possible pair of participants (McGraw & Wong, 1996) across all stimuli. After this, 100 stimuli with the lowest orientation difficulty values were selected out of the entire pool of circular stimuli (orientation difficulty range, 12.97°–19.24°) to generate scene ensembles for Experiment 5.

Composing scene orientation ensembles

After validating the scene stimuli based on orientation difficulty ratings in Experiment 4, eight circular scene ensembles of set-size six were formed for Experiment 5, with each of the eight ensembles having a unique target mean orientation with a range of 0° to 360°, while avoiding cardinal directions. To make skewed distributions of angles, five bins of 40° were created around the target mean angle (e.g., if the target mean angle was 100°, five bins would be created of angles 0°–40°, 41°–80°, 81°–120°, 121°–160°, and 161°–200°). Next, for each ensemble, one stimulus was drawn from each bin, followed by an extra stimulus drawn from either the second or fourth bin, for a total of six different orientations. The six individual orientations per ensemble were generated to have a minimum skew of 0.65 and were validated via a simulation analysis to ensure that the average orientation could only be obtained by attending globally to all orientations (see Supplementary Materials and Figures). The mean orientation for each ensemble was calculated as the orientation of the mean resultant vector of all six individual orientations combined, with each orientation being represented as a unit vector.

Calculating task performance metrics

In Experiments 2, 3, and 5 (scene ensemble rating experiments), two types of Pearson correlations (log-transformed into Fisher Z values) were calculated for each combination of scene rating, set size, and presentation time (when applicable), separately for each participant. The first type of Pearson correlation—which will be referred to from this point on as the task performance metric—examined the relationship between a participant's average rating of the ensemble for each of the ensemble set-size conditions (subsets: one, two, and four items; whole set: six items), and the predicted average rating of what was presented on the screen. For scene content and spatial boundary ratings, this predicted value was formed by taking the mathematical average of the ratings of these features for individual scenes presented in Experiment 1. For orientation ratings, this predicted value was generated by taking the mathematical average of the orientation responses for individual scenes made in Experiment 4 (alpha criterion (α) equal to 0.05). For example, the average spatial boundary estimate given by a participant for a four-item ensemble would be correlated with what the mathematical average spatial boundary value should be based on ratings of the same four individual scenes made by a separate group of participants in Experiment 1. The purpose of the task performance metric was to assess how difficult it was for participants to generate average ensemble ratings, separately for each set-size. In other words, how close is the participants response for each set size to the predicted average value for that particular set size? If task performance is independent of set-size, then the correlations should remain constant across set-sizes. However, if task performance worsens with increased ensemble set size, then the correlations would significantly decrease monotonically with increased set-size. This procedure yielded four correlations for each combination of stimulus duration and scene rating for each participant ($4 \times 4 \times 2 = 32$ total correlations for each participant in Experiment 2, and $4 \times 1 \times 2 = 8$ correlations for each participant in Experiment 3, and $4 \times 1 \times 1 = 4$ correlations for each participant in Experiment 5).

Calculating ensemble integration metrics

The second type of correlation—which will be referred to from this point on as the ensemble integration metric—examined the relationship between a participant's average rating of the ensemble for each of the ensemble set-size conditions (subsets: one, two, and four items; whole set: six items), and the predicted average rating of the original six-item whole set ensemble that each subset was constructed from. The predicted average rating of the original six-item whole set was calculated by averaging the individual scene ratings collected in an independent group of participants (Experiment 1: scene content and spatial boundary ratings; Experiment 4: scene orientation reports; $\alpha = 0.05$). The purpose of the ensemble integration metric was to assess how much individual scene information is integrated into participants' ratings of average scene content, spatial boundary, and orientation. On one-item subset trials, if we assume that participants are performing the task accurately and that there is enough skew and variance within the whole-set ensemble's individual values, then we expect that one-item subset ratings would be only weakly correlated with the predicted average ratings of the entire six-item whole-set ensemble. In other words, if a participant is only presented with one out of the six scenes that makes up the whole set ensemble, and if we assume a participant needs all six items to obtain the six-item average, then they would have no way of providing an accurate rating similar to the 6-item whole set average based on seeing only one out of the six images. However, as we increase the number of items seen to two, four, and finally to six, we expect the ensemble integration metric to increase monotonically, since participants would progressively see more visual information that can be used to approximate the true value of the six-item average (Yamanashi Leib et al., 2016). This procedure yielded the same number of correlations as described for the task performance metric above (Haberman & Whitney, 2010).

Mixed model analyses

For each rating, set-size, and presentation time, mixed model analyses were run across participants ($\alpha = 0.05$), with ensemble set-size and stimulus repetition (and their interaction) predicting Fisher Z values (either the task performance metric or the ensemble integration metric), with a random intercept for subjects run, and random slopes for set-size and repetitions for each subject run. In addition, in Experiment 2, data across presentation times was collapsed, and presentation times (and their interactions) were added as additional factors for mixed model analyses predicting task performance and ensemble integration metrics.

Scene content and spatial boundary task performance relationship

For Experiments 2 and 3, to examine the relationship (i.e., independent or correlated) in task performance between scene content and spatial boundary, two types of Pearson correlations were performed. The first correlation ($\alpha = 0.05$) examined the relationship between scene content and spatial boundary ratings for single scenes. Specifically, we correlated the Fisher Z values obtained from participants' task performance metrics for scene content and spatial boundary ratings in the 1-item subset condition. The second correlation ($\alpha = 0.05$) examined the relationship between the processing of scene content and spatial boundary for full ensembles. In this analysis we correlated the Fisher Z values obtained from participants' task performance metrics for scene content and spatial boundary ratings in the six-item whole set condition. Pearson correlations had corresponding Bayes factors (BF_{10}) reported using JASP software (JASP Team, 2025) to verify nonsignificant effects where appropriate.

Outlier analysis

We assume that participants should be able to provide similar ratings for single scenes in Experiments 2 and 3 (i.e., in the 1-item subset condition) when compared to ratings of the same scenes presented individually in Experiment 1. Similarly, we expected the reports of orientation in the one-item condition in Experiment 5 to correlate with the individual reports of orientation made in Experiment 4. To explore this, we correlated the ratings of single scenes in Experiments 2 and 3 with the same ratings in Experiment 1, and the reports of single orientation in Experiment 5 with those made in Experiment 4, and excluded participants that had nonsignificant correlations (i.e., a p value higher than 0.05). This led to five, four, and zero participants being excluded from Experiments 2, 3, and 5, respectively (resulting Ns = 50, 38, and 61). In addition, after Fisher Z values for each participant had been calculated for each set-size (either for the task performance or ensemble integration metric), Fisher Z values that were more than 1.5 times the IQR above or below the third and first quartile Fisher Z value were excluded, and this was done for each presentation time and scene feature rated. This led to slightly different sample sizes for the average Fisher Z value for a given set-size in an experiment when compared to the overall sample size for each experiment during mixed model analysis (see Experiment 2, Experiment 3, and Experiment 5 below).

Experiment 1

In Experiment 1 participants were asked to rate individual scene stimuli based on their scene content (i.e., naturalness or manufacturedness) and spatial boundary (i.e., openness and closedness), to derive reliable values for each stimulus that can then be used to calculate predicted average feature values of the ensembles presented in Experiments 2 and 3. Past research has shown participants are able to reliably ascertain differences in both scene content and spatial boundary within individual scenes (Lowe et al., 2016; Park et al., 2011), and so it is predicted that there will be high consistency in the ratings participants give for these scene features. In order to test the inter-rater reliability for the participants' ratings of the stimuli, two-way mixed intra-class correlation coefficients were calculated for the 313 stimuli for each scene feature to assess absolute agreement across participants.

Results and discussion

Excellent absolute agreement across participants was observed for ratings of both scene content (ICC = 0.99) and spatial boundary (ICC = 0.99) (Koo & Li, 2016), demonstrating that participants agreed upon their ratings of these global scene properties, and that these scene features were on average rated similarly across participants. This validated our stimulus set for the creation of scene ensembles in Experiments 2 and 3.

Experiment 2

It has been shown that people can rapidly (in under 100 ms) visually process both natural and manufactured objects (Intraub, 1981; Potter & Faulconer, 1975; Thorpe, Fize, & Marlot, 1996), individual scenes with varying amounts of naturalness and openness (Banno & Saiki, 2015; Fei-Fei, Iyer, Koch, & Perona, 2007), and object ensembles (Yamanashi Leib et al., 2016). Furthermore, superordinate categories such as scene content and openness have been shown to be processed faster than basic-level categories within scenes (Greene & Oliva, 2009a; Greene & Oliva, 2009b), suggesting that both features may be naturally diagnostic to global visual processes such as scene recognition. Moreover, summary statistics for high-level ensemble features can be extracted from groups of objects (Leib, Kosovicheva, & Whitney, 2016; Yamanashi Leib et al., 2020) and faces (Haberman & Whitney, 2007), but it is unclear if this ability extends to groups of highly complex stimuli that necessitate global visual processing (i.e., scenes). With all this in mind, Experiment 2 was conducted to determine if high-level global scene properties such as average scene content and spatial boundary can be extracted from scene ensembles, and if so, what the underlying temporal constraints might be on this ability.

We were also interested in examining how the extraction of high-level scene ensemble statistics is affected by variations in the set size of ensembles. Traditionally, ensemble perception has been shown to remain stable with increasing set-sizes (Alvarez, 2011), but most studies use simpler stimuli (e.g., geometric shapes, objects), and not more complex stimuli such as groups of scenes. If the previous result is replicated with scene ensembles, then we expect to see stable task performance for ratings of average scene content and spatial boundary as set size increases. Alternatively, if these computations on complex visual stimuli are sensitive to manipulations in set size, then we might see a decrease in task performance as set sizes increases. It has also been shown that increased stimulus presentation time can improve visual processing of both object ensembles and scenes (Fei-Fei et al., 2007; Yamanashi Leib et al., 2016). Thus we predicted that task performance metrics would increase as the stimulus presentation time increased.

Another important aspect to consider is how many items participants are globally integrating into their ensemble percepts of average scene content and spatial boundary, which speaks to the sub-sampling debate in the ensemble perception literature (i.e., can reliable summary statistics be derived by incorporating only a sub-sample of all available items in the display? Whitney & Yamanashi Leib, 2018). To investigate this, we examined how ensemble integration metric values changed with increasing set-sizes. If participants are only sampling one or two of all available scenes, then we should see a flatline trend of ensemble integration values across set sizes. However, if participants incorporate more scene information as it becomes available, then we should see ensemble integration values increase with increasing set-sizes.

Finally, we wanted to explore whether the processing of scene content and spatial boundary were correlated for single scenes and scene ensembles. Park et al. (2011) found that although PPA and LOC are sensitive to both the content and spatial boundary of single scenes, LOC was preferentially sensitive to scene content, and the PPA was preferentially sensitive to spatial boundaries. This would suggest that both features may be processed independently of each other. In contrast, Zhang, Houpt, and Harel (2019) observed that increasing naturalness in single scenes was correlated with increased openness, and increasing manufacturedness was correlated with increased closedness. This would suggest that both scene features have a more interactive relationship. We designed the scene ensembles in Experiments 2 and 3 to have no mathematical correlation between average scene content and average spatial boundary (based on Experiment 1 values), so we would be able to observe any interactivity between scene feature summary statistics if it does indeed exist.

Results and discussion

Task performance metric

The task performance metrics for all set-sizes and across all presentation times were significant for both average scene content (All $z > 0.60$, $r'_z > 0.54$, $p < 0.001$, $N > 42$; see Figure 4) and spatial boundary ratings (All $z > 0.68$, $r'_z > 0.59$, $p < 0.001$, $N > 42$; see Figure 5). This suggests that participants were able to accurately rate the scene ensembles' average features at all set-sizes regardless of stimulus presentation time.

In addition to evaluating the task performance at each set-size for a specific presentation time, we also wanted to evaluate any trends in task performance across set-sizes for a given presentation time. The purpose of this was to determine if task performance was consistently high across set-sizes, or if it declined (while still remaining accurate) with increasing set-sizes. When performing mixed model analyses on scene content task performance metrics for each presentation time, set-size was significant at all presentation times (all $t(\sim176.75) > 7.08$, $p < 0.001$, $\beta < -0.54$), repetition was significant at all presentation times ($t(\sim176.75) > 3.13$, $p < 0.001$, $\beta > 0.76$), and the set-size x repetition interaction was significant at most presentation times (for 125, 250, and 500 ms, all $t(\sim175.67) > 3.06$, $p < 0.001$, $\beta < -0.61$; see Figure 4). The set-size x repetition interaction at 1 s was nonsignificant ($t(180) = 0.48$, $p = 0.63$, $\beta = 0.09$). The effects of set-size on ratings of average scene content can be visualized by inspecting the blue regression lines with negative slope at each presentation time in Figure 4.

When performing mixed model analyses on spatial boundary task performance metrics for each presentation time, set-size was significant at all presentation times (All $t(\sim175) > 5.39$, $p < 0.001$, $\beta < -0.39$), repetition was significant at all presentation

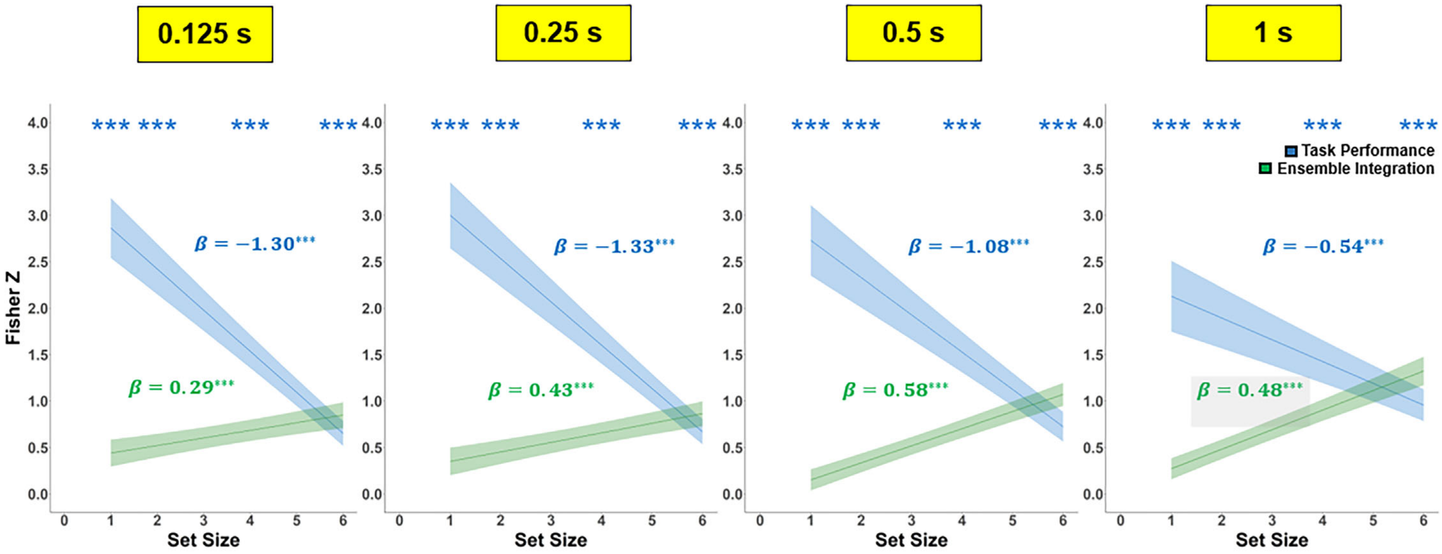

Figure 4. Scene content results for Experiment 2. Fisher *Z* values are plotted against set-size for ratings of the average scene content (i.e., naturalness/manufacturedness) of scene ensembles at 0.125, 0.250, 0.5, and 1 second (from left to right, respectively), for both the ensemble integration (green) and task performance metrics (blue). Standardized beta values for set-size (β) are provided and color-coded within each graph, and are visualized as a regression line. ***$p < 0.001$.

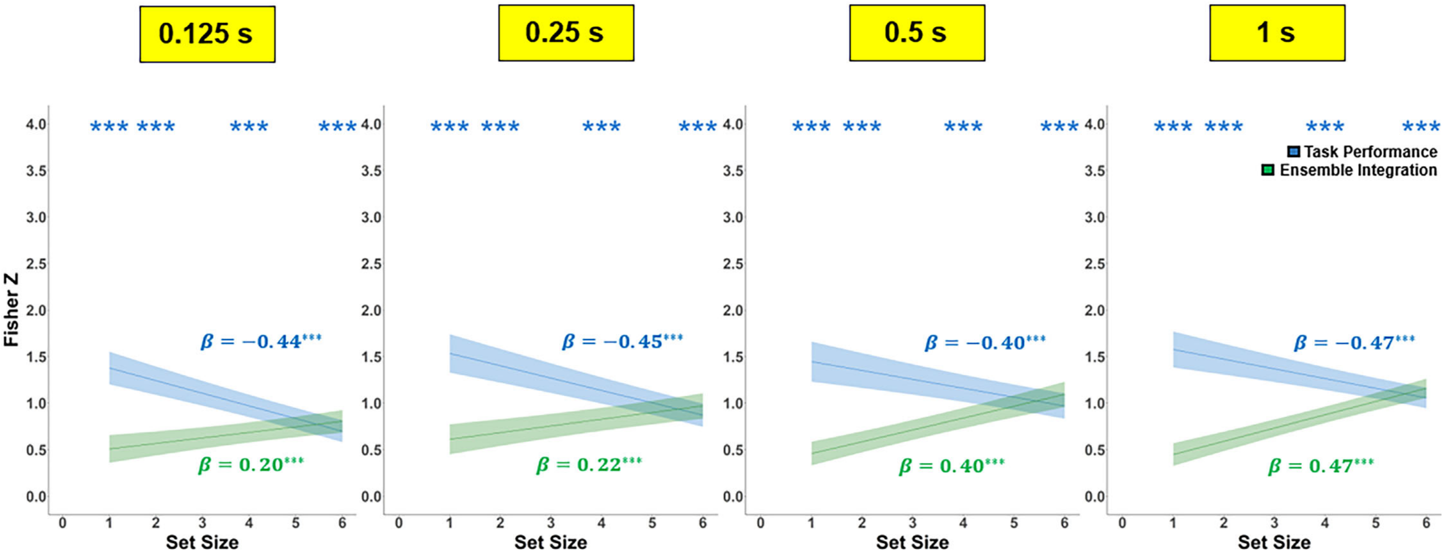

Figure 5. Spatial boundary results for Experiment 2. Fisher *Z* values are plotted against set-size for ratings of the average spatial boundary (i.e., openness/closedness) of scene ensembles at 0.125, 0.250, 0.5, and 1 second (from left to right, respectively), for both the ensemble integration (green) and task performance metrics (blue). Standardized beta values for set-size (β) are provided and color-coded within each graph, and are visualized as a regression line. ***$p < 0.001$.

times ($t(\sim 175) > 5.89$, $p < 0.001$, $\beta > 1.16$), and the set-size by repetition interaction was significant at most presentation times (for 250 ms, 500 ms and 1s, all $t(\sim 175.33) > 2.23$, $p < 0.001$, $\beta < -0.38$; see Figure 5). The set-size x repetition interaction at 125 ms was nonsignificant ($t(174) = 1.33$, $p = 0.18$, $\beta = -0.17$). As above, the effects of set-size on ratings of average spatial boundary ratings can be visualized by inspecting the blue regression lines with negative slope at each presentation time in Figure 5. Together, these results suggest that for ratings of both average scene content and spatial boundary, increasing set-size led to a decrease in task performance. This contrasts with typical findings in the ensemble literature, which demonstrate that task performance remains stable with increasing set-sizes (e.g., Alvarez, 2011). This is also consistent with literature suggesting that scene processing is more computationally demanding than object processing (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Greene & Oliva, 2009a; Oliva & Torralba, 2006), leading to a greater impact by capacity limits on its processing (Alexander et al., 2023; Kamps et al., 2016; Park et al., 2010).

In addition to looking at trends in task performance at each presentation time separately, we also wanted to evaluate if differences in presentation time itself led to changes in task performance. Although there is extensive literature suggesting that both ensemble processing and single-scene processing can occur quite efficiently at very rapid presentation times (Fei-Fei et al., 2007; Whitney & Yamanashi Leib, 2018), it is unclear whether this will remain true for scene ensembles, particularly given the results discussed above (i.e., decreasing task performance with increasing set sizes). After conducting a mixed model analysis for scene content ratings, and collapsing data across all presentation times, there were significant effects of set-size ($t(715) = 16.67$, $p < 0.001$, $\beta = -1.00$), repetition ($t(715) = 9.46$, $p < 0.001$, $\beta = 1.79$), the set-size x presentation time interaction ($t(715) = 4.05$, $p < 0.001$, $\beta = 0.24$), the set-size x repetition interaction ($t(715) = 4.87$, $p < 0.001$, $\beta = -0.72$), the presentation time × repetition interaction ($t(715) = 3.24$, $p < 0.01$, $\beta = -0.61$), and the set-size × presentation time × repetition interaction ($t(715) = 3.63$, $p < 0.001$, $\beta = 0.54$). The effect of presentation time was nonsignificant ($t(715) = 1.53$ $p = 0.13$, $\beta = -0.18$). A similar analysis on average spatial boundary ratings (after collapsing data across all presentation times) revealed significant effects of set-size ($t(708) = 7.24$, $p < 0.001$, $\beta = -0.44$), repetition ($t(708) = 7.31$, $p < 0.001$, $\beta = 1.41$), and the set-size x repetition interaction ($t(708) = 2.72$, $p < 0.001$, $\beta = -0.41$). All other variables and interactions in this analysis were nonsignificant (all $t(708) < 1.22$, $p > 0.22$, $\beta < 0.15$).

The set-size × presentation time interaction effect for scene content is visualized in Figure 4, noted as the change in the slopes of the blue task performance metric regression lines across presentation times. For scene content ratings, these slopes become significantly less negative with increasing presentation time. In contrast, the spatial boundary task performance slopes do not significantly change across presentation times (see Figure 5). Taken together, this suggests that for scene content ratings but not spatial boundary ratings, increasing presentation times seems to enhance task performance for larger set-sizes and worsens performance for smaller set-sizes. This aligns with Park and colleagues' (2011) findings that scene content and spatial boundaries are processed distinctly within the visual system, despite both being extracted via a shared mechanism that prioritizes low-level diagnostic features guided by high-level expectations during scene analysis (Bar, 2004; Lowe et al., 2016; Oliva & Schyns, 1997).

Ensemble integration metric

At each presentation time, we also wanted to verify that the ensemble whole-set average was being extracted by globally attending to all six items, and not just a subset of items. When performing mixed model analyses on scene content ensemble integration metrics for each presentation time, set-size was significant at all presentation times (all $t(\sim170) > 4.95$, $p < 0.001$, $\beta > 0.29$), repetition was significant at the 500 ms presentation time ($t(170) = 3.89$, $p < 0.001$, $\beta = -0.41$), and the set-size × repetition interaction was significant at all presentation times (all $t(\sim170) > 3.26$, $p < 0 - 0.01$, $\beta > 0.37$; see Figure 4). All other effects of repetition across presentation times were nonsignificant (all $t(\sim170) < 0.93$, $p > 0.35$, $\beta < 0.14$). The effects of set-size on ratings of average scene content can be visualized by inspecting the green regression lines with positive slopes at each presentation time in Figure 4.

When performing mixed model analyses on spatial boundary ensemble integration metrics for each presentation time, set-size was significant at all presentation times (all $t(\sim173.25) > 3.77$, $p < 0.001$, $\beta > 0.20$), repetition was significant at the 125 and 250 ms presentation times (all $t(\sim176.5) > 2.54$, $p < 0.001$, $\beta > 0.37$), and the set-size × repetition interaction was significant at 125 and 500 ms presentation times (all $t(\sim174) > 2.52$, $p < 0.001$, $\beta > 0.26$; see Figure 5). All other effects for repetition and the set-size x repetition interaction across presentation times were nonsignificant (all $t(\sim172.33) < 1.71$, $p > 0.09$, $\beta < 0.21$). The effects of set-size on ratings of average spatial boundary ratings can be visualized by inspecting the green regression lines with positive slopes at each presentation time in Figure 5. Together, the ensemble integration metric results for both scene features are consistent with the findings of Yamanashi Leib and colleagues (2016) for ensemble lifelikeness ratings, and thus suggest that participants were able to extract average scene content and spatial boundary information from scene ensembles under narrow time constraints as low as 125 ms and did so by globally integrating all information available to them as opposed to sub-sampling only one or two scenes.

As was done in the task performance analysis, we wanted to evaluate if increased presentation times affected one's ability to integrate all 6 items globally. After conducting a mixed model analysis for scene content ratings and collapsing data across all presentation times, there were significant effects of set-size ($t(688) = 6.81$, $p < 0.001$, $\beta = 0.43$), the set-size × presentation interaction ($t(688) = 2.30$, $p < 0.05$, $\beta = 0.14$), and the set-size × repetition interaction ($t(688) = 3.56$, $p < 0.001$, $\beta = 0.54$). All other variables and interactions were nonsignificant (all $t(688) < 0.92$, $p > 0.36$, $\beta > -0.12$). Similarly, a mixed model analysis for spatial boundary ratings, after collapsing data across all presentation times, showed that there was a significant effect of set-size ($t(701) = 5.56$, $p < 0.001$, $\beta = 0.30$) and the set-size x presentation time interaction was also significant ($t(701) = 2.06$, $p < 0.05$, $\beta = 0.11$). All other variables and interactions were nonsignificant (all $t(701) < 1.91$, $p > 0.06$, $\beta < 0.33$). The set-size × presentation time interaction for scene content and spatial boundary ratings is visualized in Figures 4 and 5, respectively, as the change in slope of the green ensemble integration metric regression lines across presentation times. These results suggest that for both scene features, as presentation time increases, the slopes of the ensemble integration metrics become more positive, indicating that the ensembles were increasingly processed in a more global manner. This is consistent with previous work demonstrating that a longer interval of information accumulation has a positive effect on both scene and ensemble processing (Fei-Fei et al., 2007; Roberts, Cant, & Nestor, 2019; Yamanashi Leib et al., 2016).

Correlations between scene features

There is conflicting evidence as to whether the processing of scene content and spatial boundary in individual scenes is done independently of each other (Park et al., 2011; Zhang et al., 2019). In addition, we wanted to investigate whether average scene content and spatial boundary were global features that could be independently extracted from scene ensembles. Based on individual scene content and spatial boundary values from Experiment 1, we generated scene ensembles that were uncorrelated with each other in terms of their mathematical average scene content and average spatial boundary values (See General Methods—Data Analysis). However, this mathematical independence between average scene content and spatial boundary values does not necessarily imply that we would observe perceptual independence in observers' responses. To address both of these issues, at each presentation time, Pearson correlations between scene content and spatial boundary ratings for both 1-item and 6-item presentations were calculated.

When correlating scene content and spatial boundary ratings for single scenes, stimuli presented at 125 ms ($r(35) = 0.39$, $p < 0.05$, $BF_{10} = 4.44$) and 250 ms ($r(38) = 0.35$, $p < 0.01$, $BF_{10} = 2.61$) showed significant correlations, suggesting substantial and anecdotal evidence for the alternative hypothesis, respectively, based on a Bayesian analysis (Jeffreys, 1961; see Figure 6). Correlations between scene content and spatial boundary ratings for single scenes presented for 500 ms and one second were nonsignificant (All $r(\sim42) < 0.24$, $p > 0.13$). When correlating scene content and spatial boundary ratings for six-item scene ensembles, stimuli presented at 250 ms showed a significant correlation ($r(40) = 0.38$, $p < 0.05$, $BF_{10} = 3.98$), with the Bayes Factor suggesting substantial evidence for the alternative hypothesis (Jeffreys, 1961; see Figure 6). In contrast, correlations between scene features for 6-item ensembles presented at 125 ms, 500 ms and 1 s were all nonsignificant (All $r(\sim42.66) < 0.21$, $p > 0.19$).

These results demonstrate that at shorter presentation times, scene content ratings were correlated with spatial boundary ratings of single scenes (i.e., increasing naturalness was associated with increased openness, and increasing manufacturedness was associated with increased closedness). This is consistent with the findings of Zhang and colleagues (2019), who also observed a correlation between scene content and spatial boundary ratings for single scenes. In contrast, the six-item scene ensembles, which were generated to have no correlation between average scene content and spatial boundary values across ensemble stimuli (see General Methods—Data Analysis), did not show a consistent correlation between observers' percepts of average scene content and spatial boundary across most presentation times. This indicates that summary statistics for average scene content and spatial boundary can be perceptually extracted independently of each other, even if there are associations between the processing of both features in single scenes. The one exception to this was the significant correlation at 250 ms for scene ensemble ratings, but since this correlation does not replicate in Experiment 3 (see below), this result was likely a type 1 error. Together, these results are consistent with the idea that there are separate underlying cognitive mechanisms mediating the processing of single items versus ensembles of multiple items (Cant et al., 2015).

Experiment 3

Ensemble perception can occur at speeds as fast as 100 ms, without the need to recognize individual items, for both low level (Ariely, 2001; Chong & Treisman, 2003; Dakin & Watt, 1997; Parkes et al., 2001) and high level (Haberman & Whitney, 2009; Li et al., 2016; Yamanashi Leib et al., 2016) visual features. Similarly, gist perception of individual scenes can occur just as rapidly at various levels of feature complexity (Oliva, 2005; Potter & Faulconer, 1975). The speed that these cognitive processes can occur is beyond the temporal limits of attention in object processing

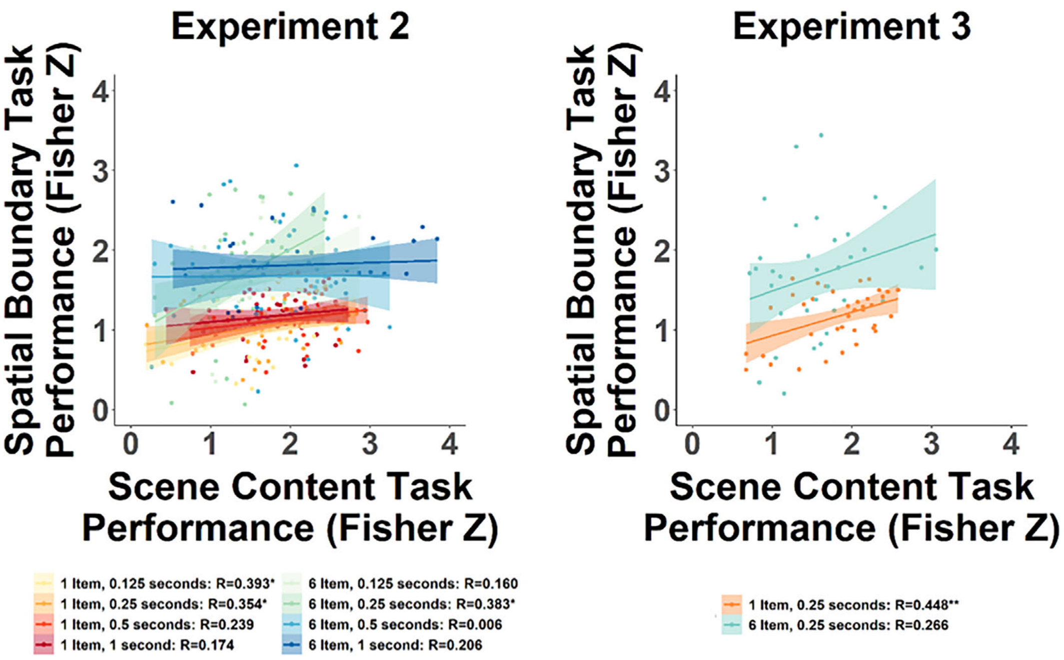

Figure 6. Correlating the processing of spatial boundary and scene content in Experiments 2 and 3. For both single scenes and six-item scene ensembles, the correlations between spatial boundary and scene content task performance metrics (Fisher *Z* values) were analyzed at all presentation times. $^*p < 0.05$, $^{**}p < 0.01$.

(Verstraten, Cavanagh, & Labianca, 2000), suggesting that ensemble processing occurs before consolidation of items into visual working memory (McNair et al., 2017; Whitney & Yamanashi Leib, 2018). However, other studies have shown that visual working memory can influence both scene (Cronin, Peacock, & Henderson, 2020) and ensemble perception (Williams et al., 2021), suggesting that these cognitive processes are not completely independent. Given these contradictory findings, the purpose of Experiment 3 was to investigate whether or not the rapid scene ensemble perception observed in Experiment 2 could be explained by the utilization of visual working memory resources. We did this by replicating Experiment 2 (at a presentation time of 250 ms to provide the potential for working memory engagement; Friedman, Cycowicz, & Gaeta, 2001; Liu, Yin, Guo, & Ye, 2024), but importantly included an additional 2AFC task to measure working memory capacity in terms of how many items were remembered in a six-item ensemble (see General Methods).

Results and discussion

Task performance metric

The task performance findings found in Experiment 3 replicated the findings in Experiment 2. Specifically, the task performance metrics for all set-sizes were significant for both average scene content (All $z > 0.61$, $r'_z > 0.54$, $p < 0.001$, $N = 38$) and spatial boundary ratings (All $z > 0.86$, $r'_z > 0.70$, $p < 0.001$, $N = 38$; see Figure 7). When performing mixed model analyses on task performance metrics for each scene feature, set-size (scene content: $t(147) = 12.21$, $p < 0.001$, $\beta = -1.21$; spatial boundary: $t(146) = 6.10$, $p < 0.001$, $\beta = -0.44$), repetition (scene content: $t(147) = 6.52$, $p < 0.001$, $\beta = 1.94$; spatial boundary: $t(146) = 5.96$, $p < 0.001$, $\beta = 1.14$) and the set-size x repetition interaction for scene content($t(147) = 3.21$, $p < 0.001$, $\beta = -0.73$) were all significant. The set-size x repetition interaction for spatial boundary was nonsignificant ($t(688) = 1.46$, $p = 0.14$, $\beta = -0.20$). Once again, the trend of decreasing Fisher *Z* values as set size increases (blue regression line, see Figure 7) suggests that although participants were able to accurately perceive the average features of the scene ensembles at all set-sizes, they had increasing difficulty with this task as more items were introduced into the ensembles.

Ensemble integration metric

The ensemble integration metric findings in Experiment 3 also replicated the results of Experiment 2. When performing mixed model analyses on ensemble integration metrics for each scene feature, significant effects were found for set-size (scene content: $t(144) = 8.20$, $p < 0.001$, $\beta = 0.49$; spatial boundary: $t(144) = 6.40$, $p < 0.001$, $\beta = 0.34$), repetition for scene

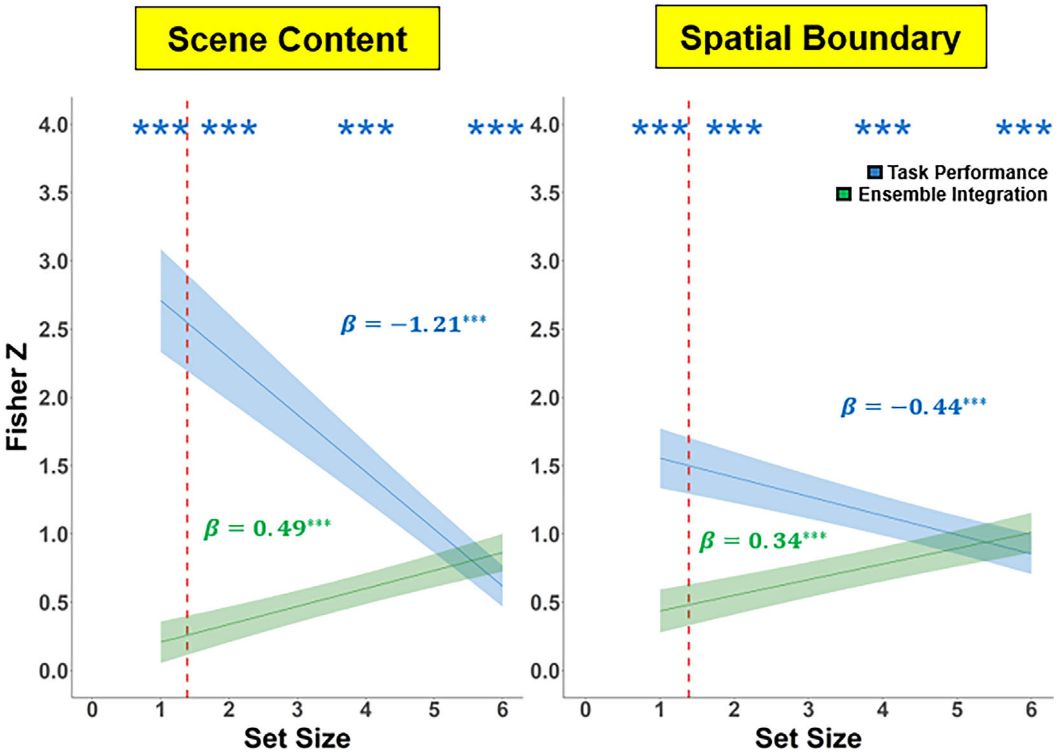

Figure 7. Results for Experiment 3. Fisher *Z* values are plotted as a function of set-size for scene content (left) and spatial boundary (right) ratings of scene ensembles at 0.250 second, for both the ensemble integration (green) and task performance (blue) metrics. Standardized beta values for set-size (β) are provided and color-coded within each graph, and are visualized as a regression line. Working memory capacity for encoding ensembles made up of six scenes is visualized by a vertical red dashed line. ***$p < 0.001$.

content ($t(144) = 2.63$, $p < 0.001$, $\beta = -0.42$), and the repetition x set-size interaction for both features (scene content: $t(144) = 7.20$, $p < 0.001$, $\beta = 0.85$; spatial boundary: $t(144) = 3.63$, $p < 0.001$, $\beta = 0.36$) (see Figure 7). The effect of repetition for spatial boundary was nonsignificant ($t(144) = 0.81$, $p = 0.42$, $\beta = 0.11$). Similar to Experiment 2, participants were globally attending to and integrating all items at each set size to extract scene ensemble summary statistics. This is signaled by the increasing Fisher *Z* values as set size increases for ratings of both average scene content and spatial boundary (green regression line, see Figure 7).

Working memory capacity

We found that the average working memory capacity was 1.39 out of six scenes, suggesting that participants were not relying on working memory to extract summary statistics for both global scene features, since they could reliably integrate six scenes into their ensemble percepts (see Figure 7). This is consistent with findings demonstrating that rapid ensemble perception can occur without consolidation of individual items within visual working memory (McNair et al., 2017; Whitney & Yamanashi Leib, 2018). Based on this, we suggest that ensemble perception for high-level scene features occurs independently from working memory mechanisms.

Correlations between scene features

Similar to Experiment 2, we wanted to evaluate the interactivity (or lack thereof) between the processing of scene content and spatial boundary for both single scenes and scene ensembles. In Experiment 3, there was a significant correlation between scene content and spatial boundary ratings for single scenes ($r(36) = 0.45$, $p < 0.01$, $BF_{10} = 9.28$), with substantial evidence provided for the alternative hypothesis (Jeffreys, 1961) (see Figure 6). However, there was no significant correlation between scene content and spatial boundary ratings for 6-item scene ensembles ($r(34) = 0.27$, $p = 0.11$, $BF_{10} = 0.68$), with anecdotal evidence provided for the null hypothesis (Jeffreys, 1961). Together, these findings replicate those from Experiment 2, and collectively suggest that while the processing of scene content and spatial boundary is correlated in the perception of individual scenes (at presentation times of 250 ms or faster), the extraction of summary statistics for these global scene features

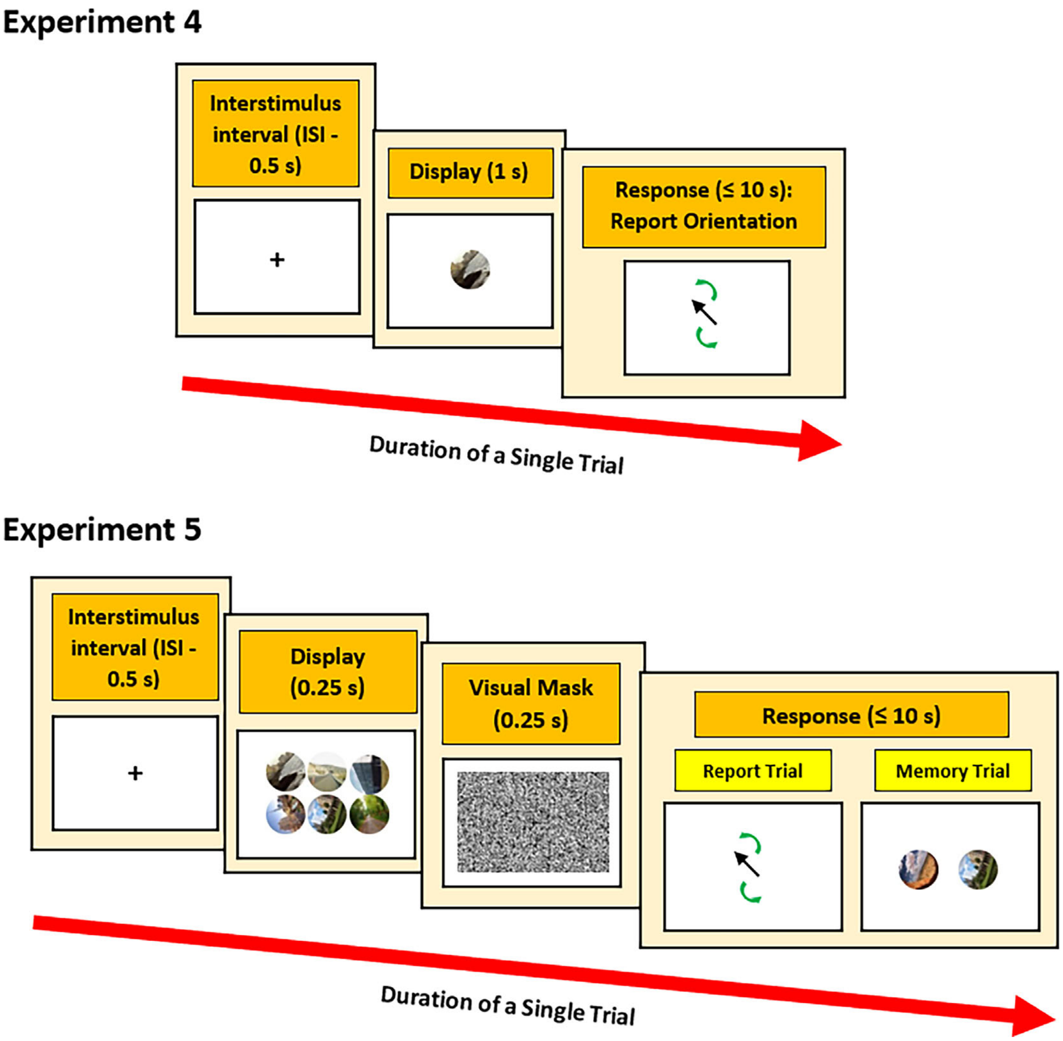
Experiment 4
Interstimulus interval (ISI - 0.5 s)
Display (1 s)
Response (≤ 10 s): Report Orientation
Duration of a Single Trial
Experiment 5
Interstimulus interval (ISI - 0.5 s)
Display (0.25 s)
Visual Mask (0.25 s)
Response (≤ 10 s)
Report Trial
Memory Trial
Duration of a Single Trial

Figure 8. Experimental design for a single reporting trial in Experiments 4 and 5, and a single memory trial in Experiment 5. At the beginning of every trial, an ISI fixation cross was presented for 0.5 second. In Experiment 4, the ISI was followed by a circular scene presented in the center of the screen for one second. For Experiment 5's reporting trials, the ISI was followed by a one-, two-, four-, or six-item scene ensemble arranged within a 512 × 768 pixel grid for 0.250 second. For Experiment 5's memory trials, the ISI was followed by 6 scene images arranged within a 512 × 768 pixel grid for 0.250 second. In Experiment 5, after all stimulus presentations, a backward mask was presented for 0.25 second. After the stimulus presentation (and masking where applicable), participants were asked to make a response within 10 seconds. For reporting trial responses in Experiments 4 and 5, participants were asked to report the orientation of a single scene, or the average orientation of the scene ensemble, respectively, using their mouse to rotate an arrow on the screen. For memory trial responses in Experiment 5, one of the six images presented was displayed, along with an image that was not a member of the previously displayed ensemble. Participants indicated which of the two images they had seen previously.

can be done independently in the perception of scene ensembles.

Experiments 4 and 5

In both Experiments 2 and 3, participants were able to rapidly extract scene ensemble summary statistics for both average scene content and spatial boundary. This was consistent with our predictions, based on the principle of feature diagnosticity. Namely, informative low-level features that are consistent with top-down scene expectations are given preferential processing, compared with less diagnostic features, to complete a given visual task (Schyns & Oliva, 1997). Interestingly, we found that task performance decreased with increasing set-sizes, and while performance was still accurate in the six-item whole-set condition, this finding suggested that participants were finding it increasingly difficult to generate summary statistics as more scenes were incorporated in the scene ensemble stimuli. This runs counter to previous findings using object and face ensembles, which collectively suggest that performance should be unaffected by increasing set sizes (Alvarez, 2011; Whitney & Yamanashi Leib, 2018). One possible explanation for this is that either the stimulus category (i.e., scenes) or the visual features extracted (i.e., scene content and spatial boundary) from scene ensembles require comparatively more higher-level visual processing resources than features extracted from object and face ensembles. Based on the design of Experiments 2 and 3, it is unclear whether the nature of the stimulus category or the computational load associated with the feature extracted was more responsible for the observed decline in task performance as set size increased.

To address this, as well as the role of feature diagnosticity in scene ensemble perception, we conducted two additional experiments where we had participants assess the orientation of either rotated single scenes (Experiment 4) or the average orientation of rotated scene ensembles (Experiment 5). Arguably, the informative low-level features for the perception of scene orientation (i.e., cardinal edges for upright scenes, oblique edges for rotated scenes) are visually simpler (i.e., less computationally demanding) to process than the low-level features used in the perception of scene content and spatial boundary (e.g., the LSFs and textures for natural scenes, and smooth spatial frequency gradients and sparse edge content for open scenes (Bar, 2004; Brady & Shafer-Skelton, 2017; Greene & Oliva, 2009a; Oliva et al., 2011; Walther & Shen, 2014).

Normally, we expect scenes to be upright, and thus will expect to prioritize attending to horizontal and vertical lines to obtain orientation information (Bar, 2004; Nasr & Tootell, 2012). For upright scenes, this expectation allows for informative cardinal edges to act as diagnostic features to extract scene orientation. However, by using rotated scenes, we are creating a conflict between the processing of low-level visual features (i.e., relying on oblique edges) and top-down expectations (i.e., relying on cardinal edges) when perceiving scene orientation (Girshick et al., 2011; Hubel & Wiesel, 1968; Nasr & Tootell, 2012; Shapley & Tolhurst, 1973). This task is made even more difficult by using circular apertures for rotated scenes. By using circular apertures and eliminating rectangular frame edges, we eliminate potential external orientation cues that could artificially bias perception, thereby isolating intrinsic scene features such as tilted horizons (Anderson et al., 2020; May & Zhaoping, 2016). This approach ensures that the task relies solely on internal scene statistics rather than external framing cues (Greene & Oliva, 2009a). This conflict was not apparent when perceiving average scene content and spatial boundary in Experiments 2 and 3, and thus serves as another test case for the principle of feature diagnosticity in ensemble scene perception. Finally, in Experiment 5, we again examined ensemble performance against visual working memory capacity, in the same manner as investigated in Experiment 3.

In Experiment 4, we had participants report the orientation value of individual rotated scenes to determine which scenes had the most perceptually discernable orientations (i.e., the lowest "orientation difficulty" scores; see General Methods – Data Analysis). In order to test the inter-rater reliability for the participants' orientation difficulty scores, two-way mixed intra-class correlation coefficients were calculated for orientation reports of all stimuli to examine absolute agreement across average responses. Next, the "best" 100 stimuli were selected (i.e., scenes with the lowest orientation difficulty scores) to generate 8 scene ensembles to use in Experiment 5 (see General Methods—Data Analysis). If task performance when estimating average orientation still decreases with increasing set-size in Experiment 5 (similar to ratings of average scene content and spatial boundary), then it is likely the nature of the stimulus category used (i.e., scene vs. object and face ensembles) that explains the task difficulty trend observed in Experiments 2 and 3. However, if task performance does not decline with increasing set sizes when estimating average orientation, then the computational load associated with the average feature extracted (i.e., average scene content and spatial boundary vs. orientation) likely explains the task difficulty trend observed in Experiments 2 and 3.

With regard to item integration, since object orientation can be rapidly and accurately extracted from object ensembles (Dakin & Watt, 1997), one possibility is that average scene orientation will also be rapidly and globally extracted from scene ensembles (i.e., the ensemble integration metric will increase with increasing set sizes). However, there is another possible finding. The perception of average orientation from rotated scenes involves the processing of low-level oblique edges that are in conflict with high-level expectations of using cardinal edges to determine orientation during scene processing (Girshick et al., 2011; Hubel & Wiesel, 1968; Nasr & Tootell, 2012; Shapley & Tolhurst, 1973). Because of this conflict between the bottom-up processing of visual cues and top-down expectations, which was not apparent when processing average scene content and spatial boundary in Experiments 2 and 3, it may be quite challenging to extract average orientation from a group of rotated scenes (i.e., the ensemble integration metric will not increase with increasing set sizes). This lack of multiple item integration may occur despite orientation appearing to be a more basic lower-level visual feature compared to scene content and spatial boundary, thus revealing the importance of feature diagnosticity in scene ensemble processing.

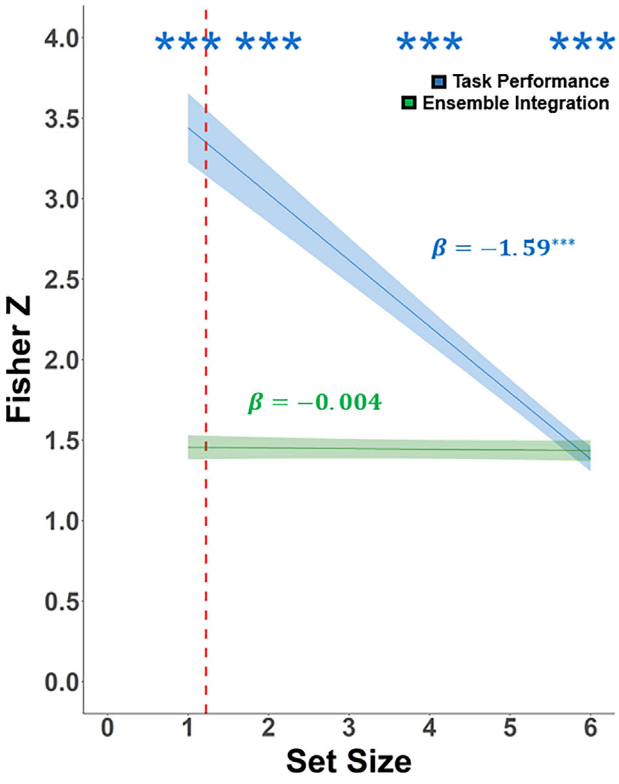

Figure 9. Results for Experiment 5. Fisher Z values are plotted as a function of Set-Size for estimates of average scene orientation at 0.250 second, for both the ensemble integration (green) and task performance (blue) metrics. Standardized beta values for set-size (β) are provided and color-coded within each graph and are visualized as a regression line. Working memory capacity for encoding ensembles made up of six scenes is visualized by a vertical red dashed line. ***$p < 0.001$.

Experiment 4 results and discussion

For Experiment 4, good absolute agreement was observed for participants' orientation difficulty scores (ICC = 0.87; Koo & Li, 2016), suggesting that people were consistent in their ability to determine the orientation of each individual scene.

Experiment 5 results and discussion

Task performance metric

First, we wanted to determine whether task performance metrics were significant across set-sizes, suggesting that perceptually reported average scene ensemble orientations were significantly correlated with the correct mathematical average orientation. This was the case, as the task performance metrics for all set-sizes were significant (All $z > 1.38$, $r'_z > 0.88$, $p < 0.001$, $N = 61$; see Figure 9). Next, we wanted to determine if task performance metrics decreased with increasing set-sizes (as was found in Experiments 2 and 3). When performing mixed model analyses on task performance metrics, set-size ($t(232) = 14.30$, $p < 0.001$, $\beta = -1.59$), repetition ($t(232) = 10.80$, $p < 0.001$, $\beta = 3.64$) and the set-size × repetition interaction ($t(232) = 7.77$, $p < 0.001$, $\beta = -2.08$) were all significant. Thus the results of Experiments 2 and 3 were replicated, but with a different summary statistic. Participants had difficulty reporting the average orientation from scene ensembles as set size increased (see the blue regression line in Figure 9). This suggests that it is the nature of the stimulus category used (i.e., scene vs. object and face ensembles) that explains the task difficulty trend observed in Experiments 2 and 3, rather than the computational load associated with the average feature extracted (i.e., average scene content and spatial boundary vs. orientation).

Ensemble integration metric

To investigate the principle of feature diagnosticity in scene ensemble processing, we examined ensemble integration metrics for participants' reports of average scene orientation. Recall that support for feature diagnosticity would be reflected as similar ensemble integration values across set sizes, likely due to a mismatch between the processing of bottom-up visual cues in rotated scenes (i.e., oblique edges) and top-down expectations of orientation information from more frequently encountered upright scenes (i.e., cardinal edges).

When performing mixed model analyses on ensemble integration metrics, the effect of repetition ($t(227) = 2.63$, $p < 0.05$, $\beta = 0.78$) was significant, but the effect of set-size ($t(227) = -0.04$, $p = .97$, $\beta = -0.004$) and the repetition x set-size interaction ($t(227) = -0.04$, $p = .97$, $\beta = -0.01$) were nonsignificant (see Figure 9). These results suggest that, unlike in Experiments 2 and 3 where participants globally integrated average scene content and spatial boundary values of scene ensembles, here participants were not able to globally integrate all 6 scenes into their estimates of average scene orientation (see the green regression line in Figure 9). Instead, participants relied on sub-sampling strategies to generate percepts of average orientation that approximated, but did not closely match, the true mathematical average of the full six-item scene ensemble. Together with the results of Experiments 2 and 3, these results provide compelling support that the principle of feature diagnosticity contributes to scene ensemble processing. That is, top-down expectations based on prior experience and stored knowledge guide the visual system to prioritize the processing of diagnostic low-level features to complete the global visual task at hand (Oliva & Schyns, 1997).

Working memory capacity

Given that average scene orientation was not globally extracted across scene ensembles, it is also likely that working memory resources were not recruited during this task. Similar to Experiments 2 and 3, we found that only 1.22 scenes were remembered out of six, suggesting that participants were not using working memory during the average orientation task.

General discussion

Across most ensemble studies, the stimuli tend to be fairly simple in visual complexity (e.g., Gabor patches, geometric shapes, objects, faces; Haberman & Whitney, 2007; Maule et al., 2014; Parkes et al., 2001). In contrast, scenes are much richer in visual information, requiring "gist" estimates of the environment (i.e., processing statistical features that correlate with scene category meaning) (Oliva, 2005). Similar to ensemble processing, gist perception helps the visual system overcome capacity limitations in attention and working memory (Dux & Marois, 2010; Luck & Vogel, 2013; Simons & Levin, 1997). One hypothesized mechanism underlying the extraction of "gist" information from single scenes is the principle of feature diagnosticity, whereby low-level visual features that are most informative for a particular high-level scene property are selectively prioritized for processing, guided by top-down expectations of the observer (Lowe et al., 2016; Oliva & Schyns, 1997; Oliva & Schyns, 2000). The purpose of the present study was to investigate the limits of ensemble processing by examining whether summary statistics could be extracted for scene ensembles. Specifically, we asked whether participants could form summary representations of average scene content, spatial boundary, and scene orientation, and also examined the role of visual working memory in these processes. Importantly, we assessed whether the formation of such global, gist-based ensemble percepts can be explained by the principle of feature diagnosticity.

The results of Experiment 1 revealed that participants could reliably rate the scene content and spatial boundary of individual scenes. Using these ratings from Experiment 1, Experiment 2 showed that participants could rapidly extract both average scene content and spatial boundary from scene ensembles at presentation times as fast as 125 ms. These results suggest that rapid coding of visual ensemble summary statistics is not limited to geometric shapes, objects or faces, but extends to visually complex scenes as well. Importantly, the ensemble average was extracted by globally attending to all six scene stimuli presented (i.e., with no subsampling of the individual items), evidenced by the rise of ensemble integration metrics as set-size increased. We argue that average scene content and spatial boundary are global ensemble properties whose low-level diagnostic features match top-down expectations, shaped by lifelong statistical learning mechanisms (Geisler, 2008). This facilitates the prioritization of these diagnostic low-level features during scene ensemble processing, and allows for the rapid extraction of summary statistics. Thus our findings demonstrate that the principle of feature diagnosticity extends from the processing of single scenes to ensembles of multiple scenes. Interestingly, although task performance remained high across set-sizes one through six when rating average scene features, it did decrease as set-size increased. This suggests that although average scene content and spatial boundary can be rapidly and accurately processed, increased perceptual and cognitive load from larger set-sizes has a modest impact on task performance. This runs counter to past object ensemble literature, which suggests that task performance remains constant even as set-size increases (Alvarez, 2011). We explored this finding further in Experiment 5 (see below).

Experiment 2 also investigated the effect of increasing presentation time on scene ensemble processing. As revealed by the sharper positive slopes for ensemble integration metrics as presentation times increased (see the green regression lines in Figures 3 and 4), the results of Experiment 2 demonstrate that the processing of average scene content and spatial boundary from scene ensembles becomes more efficient when more encoding time is available, possibly due to greater engagement of global-processing mechanisms. This is consistent with past literature suggesting that longer intervals of information accumulation have a positive effect on both scene and ensemble processing (Fei-Fei et al., 2007; Roberts et al., 2019). There was also an effect of presentation time on task performance for the processing of average scene content (but not spatial boundary), as evidenced by a more gradual decline in task performance metrics as presentation times increased (see the blue regression lines in Figures 3 and 4). This suggests that giving participants more time to extract average scene content had a beneficial impact on their ability to successfully perform the task. The fact that this trend was not observed for the processing of average spatial boundary may be explained by differences in the perceptual and neural processing of these two global scene properties. For instance, using artificially-generated scenes and fMRI multivoxel pattern analysis, Harel and colleagues (2013) showed that scene content and spatial boundary are represented distinctly, with the RSC encoding spatial boundary information, the LOC encoding content-related information, and the PPA integrating both. Furthermore, also using fMRI multivoxel pattern analysis, Kravitz, Peng, and Baker (2011) found that

PPA could reliably decode scenes based on spatial boundary (e.g., open vs. closed) but was less accurate at decoding semantic categories (e.g., naturalness, manufacturedness, place identities).

In Experiments 2 and 3 we investigated if the processing of scene content and spatial boundary were correlated, for both single scenes and scene ensembles. Past literature has suggested that these features are correlated in the processing of individual scenes, where increasing amounts of naturalness is associated with greater degrees of openness, and increasing amounts of manufacturedness is associated with greater degrees of closedness (Zhang et al., 2019). We also observed this correlation in Experiments 2 and 3, but only at shorter presentation times for single scenes (i.e., 125 ms and 250 ms). One potential reason for this is that when given only a limited amount of time to view an individual scene, the brain uses top-down mechanisms to help fill in missing information. Indeed, rapidly presented scenes undergo boundary extension when being later recalled, which is a phenomena wherein schematically consistent visual information is remembered beyond the scene's actual boundaries (Bainbridge & Baker, 2020). Given the semantic association between scene content and spatial boundary information described above, it is plausible that boundary extension mechanisms operate on both scene properties jointly to fill in missing visual information when presentation times are constrained.

Furthermore, we wanted to examine whether or not the correlation between the processing of scene content and spatial boundary for single scenes extended to the encoding of average scene content and spatial boundary in scene ensembles. The ensemble stimuli were generated such that there was no correlation between these global scene features, but we were interested in examining if this mathematical independence extended to the perceptual processing of average scene content and spatial boundary. Across presentation times, we did not observe a significant correlation between the processing of these global features from 6-item scene ensembles (except at a presentation time of 250 ms in Experiment 2, but this effect did not replicate in Experiment 3, and thus we choose not to interpret this finding as a significant effect). One explanation could be a failure of scene boundary extension mechanisms when trying to integrate multiple high-level global scene properties into a summary statistic. In addition, there could be separate cognitive mechanisms involved in the processing of single versus ensemble scene features, which would be consistent with findings from single object versus object ensemble perception (Cant et al., 2015). Taken together, the findings from Experiments 2 and 3 suggest that scene content and spatial boundary may be distinct features that can be independently extracted from scene ensembles, highlighting the brain's flexibility in processing global visual information.

The independence observed between scene content and spatial boundary ensemble processing is consistent with Park et al.'s (2011) findings that LOC and PPA are more sensitive to processing scene content and spatial boundaries, respectively (however, this was only investigated using single scenes). Moreover, OPA and RSC have also been shown to be integral in scene spatial boundary processing (Ferrara & Park, 2016; Julian, Ryan, Hamilton, & Epstein, 2016), with the OPA representing the major surfaces and planes of a scene (e.g., floors, walls, ceilings), regardless of rearrangement (Kamps et al., 2016). Future neuroimaging studies should be conducted to investigate if the independent processing of scene content and spatial boundary in scene ensembles is mediated by independent neural mechanisms in LOC and PPA/OPA/RSC, respectively, and why separate neural regions responsible for processing these global scene features does not lead to independence in the processing of single scenes.

The fact that participants can accurately extract high-level global scene ensemble features at very short presentation times suggests that selective attention may not be involved in this process. This is consistent with numerous studies arguing that attention is not necessary for ensemble perception. For example, ensemble perception of Gabor orientation has been shown to remain accurate even in instances of reduced attention (Alvarez & Oliva, 2009), and ensemble features can be accurately reported even when they are outside the focus of direct attention (Alvarez & Oliva, 2008). Furthermore, in the absence of conscious awareness, global spatial regularities found within object ensembles can increase the scope of spatial attention (Zhao & Luo, 2017). Finally, a high working memory load – which also limits resources of selective attention - does not appear to affect the accuracy of processing object ensemble summary statistics (Desimone, 1996; Downing, 2000; Epstein & Emmanouil, 2017). We contend that the principle of feature diagnosticity can explain the potential effects of attention (or lack thereof) in ensemble scene processing in the present study. That is, by prioritizing the processing of diagnostic low-level features such as spatial frequency patterns and edge structures that correspond to perceived naturalness (Geisler, 2008; Greene & Oliva, 2009a; Walther & Shen, 2014) and openness (Walther & Shen, 2014), the visual system efficiently extracts summary representations while circumventing attentional capacity limits through selective encoding and processing strategies (Hansmann-Roth, Kristjánsson, Whitney, & Chetverikov, 2021).

With this being said, findings demonstrating that attention is not required for ensemble perception do not necessarily imply that attention plays no role in ensemble processing. For example, in our study we found that task performance for average scene content ratings improved with increasing presentation time.

This is consistent with long-standing research that attention can enhance spatial resolution within an attended visual field (Yeshurun & Carrasco, 1998a; Yeshurun & Carrasco, 1998b), with longer presentation times leading to improved information accumulation (Carrasco & McElree, 2001). Moreover, within the ensemble perception literature, attended items have been shown to make heightened contributions to mean size estimates compared to non-attended items (De Fockert & Marchant, 2008). Similarly, more salient items within an ensemble tend to be weighed more in the perceived ensemble average than less salient items, without fully discounting the less salient items (Iakovlev & Utochkin, 2021). More recently, Knox and colleagues (2024) found that action-driven attention towards visual feature cues can bias later reports of average ensemble size, but this effect occurs only when attention is directed at task-relevant (i.e., size) but not irrelevant (i.e., color) features. Furthermore, when attention is efficiently captured in demanding dual-task paradigms, there can be inattentional blindness to changes in low-level object-ensemble properties (Jackson-Nielsen, Cohen, & Pitts, 2017), suggesting at least some minimum level of attention may be necessary for ensemble perception. Together, while there is not consensus on whether or not attention is required for ensemble perception, there is growing evidence in the literature supporting our findings that attention can modulate scene ensemble processing (at least for average scene content ratings).

Another key aspect of ensemble perception is that working memory resources are not required (Whitney & Yamanashi Leib, 2018), since single object recognition is not necessary to extract ensemble summary statistics (Haberman, Brady, & Alvarez, 2015). For instance, average circle size can still be extracted during object substitution masking, which greatly reduces the visibility of individual circles (Choo & Franconeri, 2010). Furthermore, complex object ensemble features like average lifelikeness and economic value do not require working memory resources to be encoded (Yamanashi Leib et al., 2016, 2020). However, recent studies have suggested that the contents of working memory can interfere with ensemble processing. Williams et al. (2021) found that reports of the average orientation of differently colored lines become skewed towards the average orientation of a subset of lines that match the color of a novel object held in working memory. Furthermore, these findings were replicated with high-level stimuli, replacing mean orientation of different colored lines with mean facial identity judgements of colored faces (Pan, Zheng, Li, & Wang, 2022). The results of Experiment 3 in the current study demonstrated that the processing of average scene content and spatial boundary did not require working memory resources, since on average less than 1.39 items were remembered out of 6, yet results from the ensemble integration analysis revealed that participants were incorporating up to 6 items into their ensemble percepts. On the surface, this seems to be at odds with the results of Williams and colleagues (2021). However, that study used a dual-task design, wherein participants first held the color and shape of a novel object in working memory, then made an ensemble judgment on a separate stimulus display, and finally made a same/different judgment based on the features of the object held in visual working memory. Our study did not utilize such a dual-task design, and found no involvement of working memory in the extraction of high-level scene ensemble features. Future studies should investigate if such a finding replicates when the contents of visual working memory are occupied with a visual feature that is correlated with a feature from the ensuing ensemble display.

In both Experiments 2 and 3, we found that increasing set-sizes led to a decline in task performance, for ratings of both average scene content and spatial boundary. These results are not consistent with previous studies demonstrating that the efficiency of ensemble perception remains stable (Alvarez, 2011; Alvarez & Oliva, 2008; Haberman & Whitney, 2007). These discrepant findings can be attributed to at least two different scenarios. On the one hand, if the use of visually complex scenes (vs. simpler geometric shapes, objects, or faces, as traditionally used in the literature) in our ensembles led to a decline in performance, we would expect task performance to still decline with increasing set-sizes for the processing of a different ensemble scene feature, namely average scene orientation. On the other hand, if the high-level nature of the scene feature being processed led to the decline in performance, then we should not observe the same results when participants extract average scene orientation, since it is a lower-level feature compared with average scene content and spatial boundary, and hence has less of a computational load associated with its processing. To investigate this, in Experiment 5 we had participants report the average orientation of scene ensembles made up of individually rotated scenes. We found that despite scene orientation being a visually simpler scene feature to process (i.e. requiring only attention to oblique edges) compared with scene content and spatial boundary, we again observed a decrease in task performance as set sizes increased, replicating the results of Experiments 2 and 3 which focused on higher-level features of scene content and spatial boundary. These results suggest that it is the visual complexity of the stimulus category that impacted ensemble task performance, as opposed to the computational load associated with the processing of the summary feature in question.

Interestingly, unlike average scene content and spatial boundary, participants had difficulty globally integrating information about average orientation across all scenes. This was likely because the low-level feature that needed to be processed to extract average orientation in our rotated scenes (i.e., oblique edges) conflicted with the high-level top-down expectations. Participants expect to utilize cardinal edges to assess scene orientation. since scenes are typically encountered upright (Girshick et al., 2011; Hubel & Wiesel, 1968; Nasr & Tootell, 2012; Shapley & Tolhurst, 1973). The consequence of this manipulation was that the visual system had difficulty prioritizing the oblique edges in rotated scenes for the rapid extraction of average orientation. Thus, instead of globally integrating orientation information from all images, it appears as though participants relied on a strategy where a smaller number of individual scenes was subsampled in an attempt to accomplish the task.

Taken together, the results of our study suggest that global visual processing circumvents limitations in attention and memory resources via the principle of feature diagnosticity. Based on findings using individual scenes, we assume this is due to preferentially processing spatial frequency information and edge statistics when encoding average scene features. For average scene content ratings, this would involve prioritizing low spatial frequencies and complex textures for natural scenes (Geisler, 2008; Greene & Oliva, 2009a) versus high spatial frequencies and cardinal edges for manmade scenes (Greene & Oliva, 2009a; Walther & Shen, 2014). For average spatial boundary ratings, this would involve prioritizing smooth spatial frequency gradients and sparse edges for open scenes (Park et al., 2011) versus abrupt high spatial frequency transitions and dense edges for closed scenes (Park et al., 2011; Walther & Shen, 2014). For average scene orientation ratings, the informative obliques edges could not act as diagnostic cues for the task, because they were in conflict with the top-down expectation of using cardinal edges, preventing scene ensemble integration (Girshick et al., 2011; Nasr & Tootell, 2012).

Alternatively, it may be that scene ensemble perception and the formation of summary statistics is achieved via the processing of other low-level features. One mechanism could be through detecting contrast differences in scenes, since improving contrast has been shown to facilitate scene recognition (Sebastian, Seemiller, & Geisler, 2020). In addition, color might act as a diagnostic feature to facilitate scene ensemble processing. For example, Gegenfurtner and Rieger (2000) have demonstrated that color enhances scene recognition by facilitating both sensory (e.g., detecting scene identity using diagnostic color cues such as the blue water of an ocean scene, or the green leaves of a forest scene) and cognitive processing (e.g., enhancing memory encoding and retrieval of scene identity through distinctive color cues). Future studies should be conducted to explicitly determine which low-level visual properties guide feature diagnosticity during scene ensemble processing. Interestingly, Kanaya, Hayashi, and Whitney (2018) found amplification effects in ensemble perception, where salient low-level visual features within the ensemble biased ensemble average percepts. Future studies could explore this further by taking diagnostic features within scenes and modulating their salience. As set-size increases, one could expect to see decreases in task performance to be accentuated with lower feature saliency, and attenuated with higher feature saliency.

In addition to relying on diagnostic low-level features, the principle of feature diagnosticity involves top-down expectations that help to prioritize certain low-level features during perceptual processing (Schyns & Oliva, 1997). This is consistent with an extensive literature highlighting how expectations influence scene processing. For example, in priming paradigms, target scenes are recognized more quickly and accurately when initially primed with scenes of the same spatial layout (Sanocki & Epstein, 1997). More recently, McLean and colleagues (2023) had participants view sequences of scenes leading to an expected destination (e.g., walking down a sidewalk leading to a store interior). They found that participants processed the gist of the final scene more quickly and accurately if it was congruent (e.g., store interior) versus incongruent (e.g., a bedroom) with their expectations. When looking at the effects of long-term semantic memory, local objects in natural scenes are more easily processed when they are placed within congruent versus incongruent scenes (Davenport & Potter, 2004), and when they are placed in typical versus atypical locations (Kaiser & Cichy, 2018). Furthermore, expectations of typical scene function (e.g., kitchens are for cooking food, bedrooms are for sleeping) have been shown to better predict scene categorization than object content and low-level visual features (Greene, Baldassano, Esteva, Beck, & Fei-Fei, 2016). These known high-evel and top-down effects on scene processing could help explain the mechanisms underlying feature diagnosticity in ensemble processing. Namely, during scene ensemble processing, expectations of which low-level scene features should be present in one's visual field can help the visual system quickly locate diagnostic low-level information for the task at hand (e.g., the formation of scene-based global summary statistics).

Overall, across 5 experiments we demonstrated that both average scene content and spatial boundary can be rapidly, accurately, and globally extracted from scene ensembles, and without reliance on working memory resources. By prioritizing the processing of diagnostic low-level features that are consistent with top-down expectations of scene identity, the visual system is able to mitigate attentional limitations during scene ensemble processing. However, unlike previous results in the ensemble literature (using simpler stimuli; see Alvarez, 2011; Alvarez & Oliva, 2009; Haberman & Whitney, 2007), scene ensemble task performance decreased with increasing set-sizes, suggesting that the visual complexity of scenes makes extraction of summary statistics more challenging compared with other ensemble stimuli. Finally, participants were not able to globally integrate multiple scenes when attempting to form percepts of average scene orientation, since information about the oblique edges required to extract average scene orientation in rotated scenes was in conflict with top-down expectations of using cardinal edges in more typically encountered upright scenes (Girshick et al., 2011; Hubel & Wiesel, 1968; Nasr & Tootell, 2012; Shapley & Tolhurst, 1973). This suggests that there are inherent perceptual limitations when extracting certain ensemble summary statistics. Together, these novel results reveal the flexibility of ensemble perception in being able to extract global features from visually complex stimuli, and highlight the importance of the principle of feature diagnosticity in this process.

Keywords: scene perception, ensemble perception, vision, working memory

Acknowledgments

Supported by an NSERC Discovery Grant (435647) to J.S.C.

Commercial relationships: none.
Corresponding author: Jonathan S. Cant.
Email: vignash.tharmaratnam@mail.utoronto.ca.
Address: University of Toronto, Toronto, ON M1C 1A4, Canada.

References

Albrecht, A. R., & Scholl, B. J. (2010). Perceptually averaging in a continuous visual world: Extracting statistical summary representations over time. *Psychological Science, 21*(4), 560–567, https://doi.org/10.1177/0956797610363543.

Alexander, A. S., Place, R., Starrett, M. J., Chrastil, E. R., & Nitz, D. A. (2023). Rethinking retrosplenial cortex: Perspectives and predictions. *Neuron, 111*(2), 150–175, https://doi.org/10.1016/j.neuron.2022.11.006.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15*(3), 122–131, https://doi.org/10.1016/j.tics.2011.01.003.

Alvarez, G. A., & Oliva, A. (2008). The representation of ensemble visual features outside the focus of attention. *Psychological Science, 19*(4), 392–398, https://doi.org/10.1167/7.9.129.

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences, 106*(18), 7345–7350, https://doi.org/10.1073/pnas.0808981106.

Alwis, Y., & Haberman, J. M. (2020). Emotional judgments of scenes are influenced by unintentional averaging. *Cognitive Research: Principles and Implications, 5*(1), https://doi.org/10.1186/s41235-020-00228-3.

Anderson, N. C., Bischof, W. F., Foulsham, T., & Kingstone, A. (2020). Turning the (virtual) world around: Patterns in saccade direction vary with picture orientation and shape in virtual reality. *Journal of Vision, 20*(8), 1–19, https://doi.org/10.1167/JOV.20.8.21.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*(2), 157–162, https://doi.org/10.1111/1467-9280.00327.

Bainbridge, W. A., & Baker, C. I. (2020). Boundaries extend and contract in scene memory depending on image properties. *Current Biology, 30*(3), 537–543.e3, https://doi.org/10.1016/j.cub.2019.12.004.

Banno, H., & Saiki, J. (2015). The processing speed of scene categorization at multiple levels of description: The superordinate advantage revisited. *Perception, 44*(3), 269–288, https://doi.org/10.1068/p7683.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5*(8), 617–629, https://doi.org/10.1038/nrn1476.

Barhorst-Cates, E. M., Rand, K. M., & Creem-Regehr, S. H. (2020). Does active learning benefit spatial memory during navigation with restricted peripheral field? *Attention, Perception, and Psychophysics, 82*(6), 3033–3047, https://doi.org/10.3758/s13414-020-02038-7.

Bastin, J., Vidal, J. R., Bouvier, S., Perrone-Bertolotti, M., Benis, D., Kahane, P., . . . Epstein, R. A. (2013). Temporal components in the parahippocampal place area revealed by human intracerebral recordings. *Journal of Neuroscience, 33*(24), 10123–10131, https://doi.org/10.1523/JNEUROSCI.4646-12.2013.

Berman, D., Golomb, J. D., & Walther, D. B. (2017). Scene content is predominantly conveyed by high spatial frequencies in scene-selective visual cortex. *PLoS ONE, 12*(12), 1–16, https://doi.org/10.1371/journal.pone.0189828.

Brady, T. F., & Shafer-Skelton, A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception* *and Performance, 43*(6), 1160–1176, https://doi.org/10.1037/xhp0000399.

Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance, 43*(6), 1160–1176, https://doi.org/10.1037/xhp0000399.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*, 433–436.

Cant, J. S., Sun, S. Z., & Xu, Y. (2015). Distinct cognitive mechanisms involved in the processing of single objects and object ensembles. *Journal of Vision, 15*(4), 12, https://doi.org/10.1167/15.4.12.

Cant, J. S., & Xu, Y. (2012). Object ensemble processing in human anterior-medial ventral visual cortex. *Journal of Neuroscience, 32*(22), 7685–7700, https://doi.org/10.1523/JNEUROSCI.3325-11.2012.

Cant, J. S., & Xu, Y. (2015). The impact of density and ratio on object-ensemble representation in human anterior-medial ventral visual cortex. *Cerebral Cortex, 25*(November), 4226–4239, https://doi.org/10.1093/cercor/bhu145.

Cant, J. S., & Xu, Y. (2017). The contribution of object shape and surface properties to object ensemble representation in anterior-medial ventral visual cortex. *Journal of Cognitive Neuroscience, 29*(2), 398–412, https://doi.org/10.1162/jocn_a_01050.

Cant, J. S., & Xu, Y. (2020). One bad apple spoils the whole bushel: The neural basis of outlier processing. *NeuroImage, 211*(October 2019), 116629, https://doi.org/10.1016/j.neuroimage.2020.116629.

Carrasco, M., & McElree, B. (2001). Covert attention accelerates the rate of visual information processing. *Proceedings of the National Academy of Sciences, 98*(9), 5363–5367, https://doi.org/10.1073/pnas.081074098.

Cate, A. D., Goodale, M. A., & Köhler, S. (2011). The role of apparent size in building- and object-specific regions of ventral visual cortex. *Brain Research, 1388*, 109–122, https://doi.org/10.1016/j.brainres.2011.02.022.

Charlton, J. A., Młynarski, W. F., Bai, Y. H., Hermundstad, A. M., & Goris, R. L. T. (2023). Environmental dynamics shape perceptual decision bias. *PLoS Computational Biology, 19*(6 June), https://doi.org/10.1371/journal.pcbi.1011104.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*(4), 393–404, https://doi.org/10.1016/S0042-6989(02)00596-5.

Choo, H., & Franconeri, S. L. (2010). Objects with reduced visibility still contribute to size averaging. *Attention, Perception, and Psychophysics, 72*(1), 86–99, https://doi.org/10.3758/APP.72.1.86.

Choo, H., & Walther, D. B. (2016). Contour junctions underlie neural representations of scene categories in high-level human visual cortex: Contour junctions underlie neural representations of scenes. *NeuroImage, 135*, 32–44, https://doi.org/10.1016/j.neuroimage.2016.04.021.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports, 6*(1), 27755, https://doi.org/10.1038/srep27755.

Corbett, J. E., & Oriet, C. (2011). The whole is indeed more than the sum of its parts: Perceptual averaging in the absence of individual item representation. *Acta Psychologica, 138*(2), 289–301, https://doi.org/10.1016/j.actpsy.2011.08.002.

Cronin, D. A., Peacock, C. E., & Henderson, J. M. (2020). Visual and verbal working memory loads interfere with scene-viewing. *Attention, Perception, and Psychophysics, 82*(6), 2814–2820, https://doi.org/10.3758/s13414-020-02076-1.

Dakin, S. C., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research, 37*(22), 3181–3192, https://doi.org/10.1016/S0042-6989(97)00133-8.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15*(8), 559–564, https://doi.org/10.1111/j.0956-7976.2004.00719.x.

De Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception and Psychophysics, 70*(5), 789–794, https://doi.org/10.3758/PP.70.5.789.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proceedings of the National Academy of Sciences of the United States of America, 93*(24), 13494–13499, https://doi.org/10.1073/pnas.93.24.13494.

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

Downing, P. E. (2000). Interactions between visual working memory and selective attention. *Psychological Science, 11*(6), 467–473, https://doi.org/10.1111/1467-9280.00290.

Dux, P. E., & Marois, R. (2010). How humans search for targets through time: A review of data and theory from the attentional blink. *Attention,* *Perception & Psychophysics, 71*(8), 1683–1700, https://doi.org/10.3758/APP.71.8.1683.

Epstein, M. L., & Emmanouil, T. A. (2017). Ensemble coding remains accurate under object and spatial visual working memory load. *Attention, Perception, and Psychophysics, 79*(7), 2088–2097, https://doi.org/10.3758/s13414-017-1353-2.

Epstein, R. A., & Baker, C. I. (2019). Scene perception in the human brain. *Annual Review of Vision Science, 5*, 373–397, https://doi.org/10.1146/annurev-vision-091718-014809.

Epstein, R., Graham, K. S., & Downing, P. E. (2003). Viewpoint-specific scene representations in human parahippocampal cortex. *Neuron, 37*(5), 865–876, https://doi.org/10.1016/S0896-6273(03)00117-X.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191, https://doi.org/10.3758/BF03193146.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision, 7*(1), 10, https://doi.org/10.1167/7.1.10.

Ferrara, K., & Park, S. (2016). Neural representation of scene boundaries. *Neuropsychologia, 89*, 180–190, https://doi.org/10.1016/j.neuropsychologia.2016.05.012.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A, 4*(12), 2379–2394, https://doi.org/10.1364/JOSAA.4.002379.

Friedman, D., Cycowicz, Y. M., & Gaeta, H. (2001). The novelty P3: an event-related brain potential (ERP) sign of the brain's evaluation of novelty. *Neuroscience & Biobehavioral Reviews, 25*(4), 355–373, https://doi.org/10.1016/S0149-7634(01)00019-7.

Gegenfurtner, K. R., & Rieger, J. (2000). Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology, 10*(13), 805–808, https://doi.org/10.1016/S0960-9822(00)00563-7.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology, 59*, 167–192, https://doi.org/10.1146/annurev.psych.58.110405.085632.

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience, 14*(7), 926–932, https://doi.org/10.1038/nn.2831.

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology: General, 145*(1), 82–94, https://doi.org/10.1037/xge0000129.

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

Greene, M. R., & Oliva, A. (2009b). The briefest of glances: The time course of natural scene understanding. *Psychological Science, 20*(4), 464–472, https://doi.org/10.1111/j.1467-9280.2009.02316.x.

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General, 144*(2), 432–446, https://doi.org/10.1037/xge0000053.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17*(17), R751–R753, https://doi.org/10.1016/j.cub.2007.06.039.

Haberman, J., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision, 9*(2009), 1–13, https://doi.org/10.1167/9.11.1.Introduction.

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, and Psychophysics, 72*(7), 1825–1838, https://doi.org/10.3758/APP.72.7.1825.

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin and Review, 18*(5), 855–859, https://doi.org/10.3758/s13423-011-0125-6.

Hansmann-Roth, S., Kristjánsson, Á., Whitney, D., & Chetverikov, A. (2021). Dissociating implicit and explicit ensemble representations reveals the limits of visual perception and the richness of behavior. *Scientific Reports, 11*(1), https://doi.org/10.1038/s41598-021-83358-y.

Harel, A., Kravitz, D. J., & Baker, C. I. (2013). Deconstructing visual scenes in cortex: Gradients of object and spatial layout information. *Cerebral Cortex, 23*(4), 947–957, https://doi.org/10.1093/cercor/bhs091.

Harel, A., Mzozoyana, M. W., Al Zoubi, H., Nador, J. D., Noesen, B. T., Lowe, M. X., . . . Cant, J. S. (2020). Artificially-generated scenes demonstrate the importance of global scene properties for scene perception. *Neuropsychologia, 141*, 107434, https://doi.org/10.1016/j.neuropsychologia.2020.107434.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology, 195*(1), 215–243, https://doi.org/10.1113/jphysiol.1968.sp008455.

Iakovlev, A. U., & Utochkin, I. S. (2021). Roles of saliency and set size in ensemble averaging. *Attention, Perception, and Psychophysics, 83*(3), 1251–1262, https://doi.org/10.3758/s13414-020-02089-w.

Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance, 7*(3), 604–610, https://doi.org/10.1037/0096-1523.7.3.604.

Jackson-Nielsen, M., Cohen, M. A., & Pitts, M. A. (2017). Perception of ensemble statistics requires attention. *Consciousness and Cognition, 48*, 149–160, https://doi.org/10.1016/j.concog.2016.11.007.

JASP Team, (2025). *JASP (Version 0.19)[Computer software],* https://jasp-stats.org/.

Jeffreys, H. (1961). *The Theory of Probability* (Third). Oxford, UK; Oxford University Press.

Julian, J. B., Ryan, J., Hamilton, R. H., & Epstein, R. A. (2016). The occipital place area is causally involved in representing environmental boundaries during navigation. *Current Biology, 26*(8), 1104–1109, https://doi.org/10.1016/j.cub.2016.02.066.

Jung, Y., & Walther, D. B. (2021). Neural representations in the prefrontal cortex are task dependent for scene attributes but not for scene categories. *Journal of Neuroscience, 41*(34), 7234–7245, https://doi.org/10.1523/JNEUROSCI.2816-20.2021.

Kaiser, D., & Cichy, R. M. (2018). Typical visual-field locations enhance processing in object-selective channels of human occipital cortex. *Journal of Neurophysiology, 120*(2), 848–853, https://doi.org/10.1152/jn.00229.2018.

Kamps, F. S., Julian, J. B., Kubilius, J., Kanwisher, N., & Dilks, D. D. (2016). The occipital place area represents the local elements of scenes. *NeuroImage, 132*(3), 417–424, https://doi.org/10.1016/j.neuroimage.2016.02.062.

Kanaya, S., Hayashi, M. J., & Whitney, D. (2018). Exaggerated groups: Amplification in ensemble coding of temporal and spatial features. *Proceedings of the Royal Society B: Biological Sciences, 285*(1879), 20172770, https://doi.org/10.1098/rspb.2017.2770.

Kersten, D. (1987). Predictability and redundancy of natural images. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 4*(12), 2395–2400.

Khayat, N., Pavlovskaya, M., & Hochstein, S. (2024). Comparing explicit and implicit ensemble perception: 3 stimulus variables and 3 presentation modes. *Attention, Perception, and Psychophysics, 86*(2), 482–502, https://doi.org/10.3758/s13414-023-02784-4.

Kinchla, R. A. (1977). The role of structural redundancy in the perception of visual targets. *Perception & Psychophysics, 22*(1), 19–30, https://doi.org/10.3758/BF03206076.

Kinchla, R. A., & Wolfe, J. M. (1979). The order of visual processing: "Top-down," "bottom-up," or "middle-out". *Perception & Psychophysics, 25*(3), 225–231, https://doi.org/10.3758/BF03202991.

Kleiner, M., Brainard, D. H., & Pelli, D. G. (2007). "What's new in Psychtoolbox-3?" *Perception, 36*(14), 1–16.

Knox, K., Pratt, J., & Cant, J. S. (2024). Examining the role of action-driven attention in ensemble processing. *Journal of Vision, 24*(6), 5, https://doi.org/10.1167/jov.24.6.5.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine, 15*(2), 155–163, https://doi.org/10.1016/j.jcm.2016.02.012.

Kravitz, D. J., Peng, C. S., & Baker, C. I. (2011). Real-world scene representations in high-level visual cortex: It's the spaces more than the places. *Journal of Neuroscience, 31*(20), 7322–7333, https://doi.org/10.1523/JNEUROSCI.4588-10.2011.

Leib, A. Y., Kosovicheva, A., & Whitney, D. (2016). Fast ensemble representations for abstract visual impressions. *Nature Communications, 7*, 1–10, https://doi.org/10.1038/ncomms13186.

Li, H., Ji, L., Tong, K., Ren, N., Chen, W., Liu, C. H., . . . Fu, X. (2016). Processing of individual items during ensemble coding of facial expressions. *Frontiers in Psychology, 7*(SEP), 1–11, https://doi.org/10.3389/fpsyg.2016.01332.

Liu, Q., Yin, X., Guo, L., & Ye, C. (2024). Influence of presentation duration on filtering of irrelevant stimuli in visual working memory. *BMC Psychology, 12*(1), 469, https://doi.org/10.1186/s40359-024-01969-2.

Lowe, M. X., Gallivan, J. P., Ferber, S., & Cant, J. S. (2016). Feature diagnosticity and task context shape activity in human scene-selective cortex. *NeuroImage, 125*, 681–692, https://doi.org/10.1016/j.neuroimage.2015.10.089.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends Cogn Sci.*, *17*(8), 391–400, https://doi.org/10.1016/j.tics.2013.06.006.Visual.

Maule, J., Witzel, C., & Franklin, A. (2014). Getting the gist of multiple hues: metric and categorical effects on ensemble perception of hue. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 31*(4), A93–102, https://doi.org/10.1364/JOSAA.31.000A93.

May, K. A., & Zhaoping, L. (2016). Efficient coding theory predicts a tilt aftereffect from viewing untilted patterns. *Current Biology, 26*(12), 1571–1576, https://doi.org/10.1016/j.cub.2016.04.037.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46, https://doi.org/10.1037/1082-989X.1.1.30.

McLean, D., Nuthmann, A., Renoult, L., & Malcolm, G. L. (2023). Expectation-based gist facilitation: Rapid scene understanding and the role of top-down information. *Journal of Experimental Psychology: General, 152*(7), 1907–1936, https://doi.org/10.1037/xge0001363.

McNair, N. A., Goodbourn, P. T., Shone, L. T., & Harris, I. M. (2017). Summary statistics in the attentional blink. *Attention, Perception, and Psychophysics, 79*(1), 100–116, https://doi.org/10.3758/s13414-016-1216-2.

Miller, A. P., Vedder, L. C., Law, M. L., & Smith, D. M. (2014). Cues, context, and long-term memory: The role of the retrosplenial cortex in spatial cognition. *Frontiers in Human Neuroscience, 8*, 586, https://doi.org/10.3389/fnhum.2014.00586.

Nasr, S., & Tootell, R. B. H. (2012). A cardinal orientation bias in scene-selective visual cortex. *Journal of Neuroscience, 32*(43), 14921–14926, https://doi.org/10.1523/JNEUROSCI.2036-12.2012.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology, 9*(3), 353–383, https://doi.org/10.1016/0010-0285(77)90012-3.

Oliva, A. (2005). Gist of the scene. In L. Itti, G. Rees & J.K. Tsotsos (Eds.), *Neurobiology of Attention* (pp. 251–256). New York: Academic Press, https://doi.org/10.1016/B978-012375731-9/50045-8.

Oliva, A., Park, S., & Konkle, T. (2011). Representing, perceiving, and remembering the shape of visual space. In L. R. Harris & M. R. M. Jenkin (Eds.), *Vision in 3D Environments* (pp. 308–340). Cambridge, UK: Cambridge University Press, https://doi.org/10.1017/CBO9780511736261.014.

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology, 34*(1), 72–107, https://doi.org/10.1006/cogp.1997.0667.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology, 41*(2), 176–210, https://doi.org/10.1006/cogp.1999.0728.

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, https://doi.org/10.1016/S0079-6123(06)55002-2.

Palmer, J. (1990). Attentional limits on the perception and memory of visual information. *Journal of Experimental Psychology: Human Perception and Performance, 16*(2), 332–350, https://doi.org/10.1037/0096-1523.16.2.332.

Pan, T., Zheng, Z., Li, F., & Wang, J. (2022). Memory matching features bias the ensemble perception of facial identity. *Frontiers in Psychology, 13*, 1053358, https://doi.org/10.3389/fpsyg.2022.1053358.

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

Park, S., Chun, M. M., & Johnson, M. K. (2010). Refreshing and integrating visual scenes in scene-selective cortex. *Journal of Cognitive Neuroscience, 22*(12), 2813–2822.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*(7), 739–744, https://doi.org/10.1038/89532.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442.

Peyrin, C., Baciu, M., Segebarth, C., & Marendaz, C. (2004). Cerebral regions and hemispheric specialization for processing spatial frequencies during natural scene recognition. An event-related fMRI study. *NeuroImage, 23*(2), 698–707, https://doi.org/10.1016/j.neuroimage.2004.06.020.

Potter, M. C., & Faulconer, B. A. (1975). Time to understand pictures and words. *Nature, 253*(5491), 437–438.

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, https://www.r-project.org.

Rajimehr, R., Devaney, K. J., Bilenko, N. Y., Young, J. C., & Tootell, R. B. H. (2011). The "parahippocampal place area" responds preferentially to high spatial frequencies in humans and monkeys. *PLoS Biology, 9*(4), https://doi.org/10.1371/journal.pbio.1000608.

Roberts, T., Cant, J. S., & Nestor, A. (2019). Elucidating the neural representation and the processing dynamics of face ensembles. *Journal of Neuroscience, 39*(39), 7737–7747, https://doi.org/10.1523/JNEUROSCI.0471-19.2019.

Sanocki, T., & Epstein, W. (1997). Priming Spatial Layout of Scenes. *Psychological Science, 8*(5), 374–378, https://doi.org/10.1111/j.1467-9280.1997.tb00428.x.

Schyns, P. G., & Oliva, A. (1997). Flexible, diagnosticity-driven, rather than fixed, perceptually determined scale selection in scene and face recognition. *Perception, 26*(8), 1027–1038, https://doi.org/10.1068/p261027.

Sebastian, S., Seemiller, E. S., & Geisler, W. S. (2020). Local reliability weighting explains identification of partially masked objects in natural images. *Proceedings of the National Academy of Sciences, 117*(47), 29363–29370, https://doi.org/10.1073/pnas.1912331117.

Shapley, R. M., & Tolhurst, D. J. (1973). Edge detectors in human vision. *The Journal of Physiology, 229*(1), 165–183, https://doi.org/10.1113/jphysiol.1973.sp010133.

Shikauchi, Y., & Ishii, S. (2016). Robust encoding of scene anticipation during human spatial navigation. *Scientific Reports, 6*(1), 37599, https://doi.org/10.1038/srep37599.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences, 1*(7), 261–267, https://doi.org/10.1016/S1364-6613(97)01080-2.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520.

Tiurina, N. A., Markov, Y. A., Whitney, D., & Pascucci, D. (2024). The functional role of spatial anisotropies in ensemble perception. *BMC Biology, 22*(1), 28, https://doi.org/10.1186/s12915-024-01822-3.

Verstraten, F. A. J., Cavanagh, P., & Labianca, A. T. (2000). Limits of attentive tracking reveal temporal properties of attention. *Vision Research, 40*(26), 3651–3664, https://doi.org/10.1016/S0042-6989(00)00213-3.

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *Journal of Neuroscience, 29*(34), 10573–10581, https://doi.org/10.1523/JNEUROSCI.0559-09.2009.

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

Watamaniuk, S. N. J., & McKee, S. P. (1998). Simultaneous Encoding of Direction at a Local and Global Scale. *Perception & Psychophysics, 60*(2), 191.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology, 69*(1), 105–129, https://doi.org/10.1146/annurev-psych-010416-044232.

Williams, R. S., Pratt, J., Ferber, S., & Cant, J. S. (2021). Tuning the ensemble: Incidental skewing of the perceptual average through memory-driven selection. *Journal of Experimental Psychology: Human Perception and Performance, 47*(5), 648–661, https://doi.org/10.1037/xhp0000907.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 3485–3492). Piscataway, NJ: IEEE, https://doi.org/10.1109/CVPR.2010.5539970.

Yamanashi Leib, A., Chang, K., Xia, Y., Peng, A., & Whitney, D. (2020). Fleeting impressions of economic value via summary statistical representations. *Journal of Experimental Psychology: General, 149*(10), 1811–1822, https://doi.org/10.1037/xge0000745.

Yamanashi Leib, A., Kosovicheva, A., & Whitney, D. (2016). Fast ensemble representations for abstract visual impressions. *Nature Communications, 7*(13186), 1–10, https://doi.org/10.1038/ncomms13186.

Yeshurun, Y., & Carrasco, M. (1998a). Attention improves or impairs visual performance by enhancing spatial resolution. *Nature, 396*(6706), 72–75.

Yeshurun, Y., & Carrasco, M. (1998b). Spatial attention improves performance in spatial resolution tasks. *Vision Research, 39*(2), 293–306, https://doi.org/S004269899800114X.

Zeidman, P., Mullally, S. L., Schwarzkopf, D. S., & Maguire, E. A. (2012). Exploring the parahippocampal cortex response to high and low spatial frequency spaces. *NeuroReport, 23*(8), 503–507, https://doi.org/10.1097/WNR.0b013e328353766a.

Zhang, H., Houpt, J. W., & Harel, A. (2019). Establishing reference scales for scene naturalness and openness: Naturalness and openness scales.

Behavior Research Methods, 51(3), 1179–1186, https://doi.org/10.3758/s13428-018-1053-4.

Zhao, J., & Luo, Y. (2017). Statistical regularities guide the spatial scale of attention. *Attention, Perception, and Psychophysics, 79*(1), 24–30, https://doi.org/10.3758/s13414-016-1233-1.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(6), 1452–1464). Piscataway, NJ: IEEE, https://doi.org/10.1109/TPAMI.2017.2723009.