CHAPTER TWO

Scaling up visual attention and visual working memory to the real world

Timothy F. Brady*, Viola S. Störmer, Anna Shafer-Skelton, Jamal R. Williams, Angus F. Chapman, Hayden M. Schill
Department of Psychology, University of California, San Diego, La Jolla, CA, United States
*Corresponding author: e-mail address: timbrady@ucsd.edu

Contents

1. Introduction 30
2. Visual attention 31
   2.1 Introduction 31
   2.2 From simple features to multi-feature objects: Principles of attentional selection 32
   2.3 Attentional selection at the level of real-world objects and semantic categories 35
   2.4 Attention to, and processing of, "ensembles" and groups of objects 39
   2.5 Cross-modal influences on perception and attention: Real-world objects make sounds 42
   2.6 Visual attention conclusion 46
3. Visual working memory 47
   3.1 Introduction 47
   3.2 Visual working memory stores not just individual items, but ensemble information 48
   3.3 The world has more than objects: Scenes and surfaces in visual memory 52
   3.4 We have knowledge about the objects we wish to remember: Learning and knowledge in visual working memory 54
   3.5 Expertise and visual working memory 58
4. Conclusion 59
References 60

Abstract

Both visual attention and visual working memory tend to be studied either with very simple stimuli and low-level paradigms, which are designed to allow us to understand the representations and processes in detail, or with fully realistic stimuli that make such precise understanding difficult but are more representative of the real world. In this chapter we argue for an intermediate approach in which visual attention and visual

Psychology of Learning and Motivation, Volume 70
ISSN 0079-7421
https://doi.org/10.1016/bs.plm.2019.03.001

working memory are studied by scaling up from the simplest settings to more complex settings that capture some aspects of the complexity of the real-world, while still remaining in the realm of well-controlled stimuli and well-understood tasks. We believe this approach, which we have been taking in our labs, will allow a generalizable set of knowledge about visual attention and visual working memory while maintaining the rigor and control that is typical of vision science and psychophysics studies.

1. Introduction

Both visual attention and visual working memory tend to be studied with very simple stimuli and low-level paradigms. In particular, the average attention or visual working memory study presents several simple shapes (possibly in distinct colors) on a simple, homogeneous background, and then asks about the allocation of attention or the limits of working memory in such situations. There are significant benefits to this approach; in particular, using simple stimuli allows us to understand the representations of the stimuli, because we have a significant understanding of basic visual processing and of the representations of simple stimuli in the visual system; furthermore, this approach limits the ability of participants to engage in overly complicated strategies.

However, studying attention and memory with simplified stimuli, made up of discrete, single-feature objects on blank backgrounds, leaves much left unstudied about how attention and visual working memory function in more realistic settings. By contrast, studying attention and memory in fully real-world settings is challenging, and may not allow for sufficient control of the stimuli or of participants' strategies to allow researchers to fully understand the underlying representations or to make computational models that capture participants' behavior. Thus, as in many domains, the study of visual cognition faces a trade-off between deep, process-level understanding and external validity.

In this chapter we argue that both of these approaches are best supplemented and linked by an intermediate approach in which visual attention and visual working memory are studied by scaling up from the simplest settings to more complex settings that capture some aspects of the complexity of the real-world (e.g., real-world objects instead of colored squares; objects' sounds in addition to visual objects; surfaces and scenes in addition to objects), while still remaining in the realm of well-controlled stimuli and well-understood tasks. We believe this approach, which we have been taking

in our labs, can allow a more generalizable set of knowledge about visual attention and visual working memory while maintaining the rigor and control that is typical of vision science and psychophysics studies. In particular, in the current chapter we argue for scaling up visual attention (Section 2) and visual working memory (Section 3) in ways that capture certain important aspects of the real world. In particular, in the section on visual attention, we review work from our labs and others on how attention operates over objects, rather than only spatial positions or simple single features (Section 2.2); what the role of semantic categories is in selective attention (Section 2.3); the role of ensembles or sets of objects, rather than individual objects (Section 2.4); and the role of audition in visual attention and visual perception (Section 2.5). In the section on visual working memory, we review work on ensemble representations and other effects of non-independence between items (Section 3.2); on memory for surfaces and scene layout (Section 3.3); on the impact of learning and existing semantic knowledge on visual working memory (Section 3.4); and on the role of expertise in visual working memory, and what might cause the effects of expertise on capacity (Section 3.5). Together, we take the position that important work can be done using controlled stimuli and well-understood tasks, while nevertheless scaling up the study of visual attention and visual working memory to take into account more real-world factors. As we will show, this approach has afforded insights into attention and working memory that were not apparent only from the literature using simple stimuli—for example, this kind of work demonstrates the important role of semantics and high-level features in many "visual" attention and memory situations; shows the importance of considering that objects are not often treated independently, but tend to be encoded relative to each other or in sets (which allows people to circumvent capacity limits); and shows the importance of considering how multimodal stimuli like sounds, as well as visual surfaces and scene layout, are integrated into visual recognition of objects.

2. Visual attention

2.1 Introduction

Only a small portion of the information hitting our retinas reaches our conscious visual experience. At any given time, we are typically only aware of those parts of a visual scene that we are paying attention to. For example, when actively looking for something (e.g., the keys on a cluttered desk), or when being drawn to a particularly salient stimulus (e.g., the loud honk

of a car), we are not aware of all of the objects and surfaces present. Attention refers to a mechanism that selects information from the enormous number of sensory inputs and prioritizes that information for further processing. Thus, attention is fundamental for our perceptual experiences, determining what and how sensory inputs reach awareness.

Decades of research on attention have shown how attention operates in terms of spatial locations and simple features (for a review, see Carrasco, 2011). For example, by directing spatial attention to a particular region in the visual field, processing of objects appearing at that location is enhanced (Posner & Petersen, 1990). Similarly, by tuning attention to a particular feature, such as the color red, objects containing the attended color receive a boost in visual processing (Sàenz, Buraĉas, & Boynton, 2003). This line of work has been exceptionally successful at revealing basic computational principles of attentional selection, for example, by showing that attention can increase the neural gain (Luck, Chelazzi, Hillyard, & Desimone, 1997; Moran & Desimone, 1985), modulate neural noise (Mitchell, Sundberg, & Reynolds, 2007), or shift neural tuning functions (David, Hayden, Mazer, & Gallant, 2008; Motter, 1994) in sensory cortices. The majority of this research has been done using simple, low-level visual stimuli, such as oriented bars or moving dot patterns, with the benefit of tapping into a fairly strong understanding of how these simple features are represented in the visual system. However, this research does not fully capture many important aspects of how we use attention in the real world, where we need to select multi-feature objects in complex scenes, and where objects make noises and do not appear as isolated, unisensory entities.

While current theories of attention often do not explain how attentional selection occurs in more realistic scenarios, the human attention system must be equipped with mechanisms that can handle a rich, multisensory environment that is populated by complex, natural entities (e.g., animals, people, and real-world objects). What do we know about how attention operates across multi-feature objects, real-world objects and semantic categories? How does attention deal with inputs across sensory modalities? Here, we review recent papers that address some of these questions and discuss outstanding challenges in understanding how attention operates in more realistic scenarios.

2.2 From simple features to multi-feature objects: Principles of attentional selection

Many theories of visual attention highlight the importance of simple features, such as color, orientation, or direction of motion, for stimulus selection. These "feature-based" theories are supported by findings of enhanced visual

processing (Ling, Liu, & Carrasco, 2009; Liu, Stevens, & Carrasco, 2007) and increased neural activity (Liu, Larsson, & Carrasco, 2007; Liu, Slotnick, Serences, & Yantis, 2003; O'Craven, Rosen, Kwong, Treisman, & Savoy, 1997; Sàenz, Buracâs, & Boynton, 2002) when attention is directed to a relevant feature (e.g., a specific color or direction of motion) among other irrelevant features. Many studies have shown that this attentional enhancement is not spatially isolated, but acts to boost processing throughout the visual field (Andersen, Hillyard, & Müller, 2013; Sàenz et al., 2003; Serences & Boynton, 2007; Störmer & Alvarez, 2014; White & Carrasco, 2011), influencing the processing of visual information outside the spatial focus of attention (Fig. 1A). This global feature enhancement lead to the development of the "feature-similarity gain model" (Martinez-Trujillo & Treue, 2004; Treue & Martinez-Trujillo, 1999), which claims that attention modulates the responses of individual neurons based on the similarity between the tuning of that neuron and the attended feature. As a result of this modulation, attention to a specific feature will lead to increased population-level tuning toward the attended feature, and enhanced processing of that feature throughout the visual field. Similar to this feature-based selection account, simple objects, such as surfaces of spatially intermingled dots defined by color and motion direction, can be selectively attended (Ciaramitaro, Mitchell, Stoner, Reynolds, & Boynton, 2011; Ernst, Boynton, & Jazayeri, 2013; Wannig, Rodríguez, & Freiwald, 2007), even though the surfaces occupy the same spatial location. Attention to surfaces results in increased neural activity to the attended surface and improved behavioral performance (Ciaramitaro et al., 2011; Wannig et al., 2007), just like attention to simple isolated features.

How does this basic principle of global gain for simple features and surfaces relate to the selection of more complex, real-world objects? Real-world objects consist of multiple features that need to occur in a particular configuration to create a cohesive object (e.g., to perceive a face, eyes need to be aligned next to each other, and nose and mouth need to appear below to compose the canonical configuration of a face). Furthermore, as they occur in the real world, objects often differ in many simple lower-level features, such as orientation, color, or size, resulting in a substantial variety of low-level features even within an object category. Despite these difficulties, we are extremely quick in detecting objects in visual scenes (Potter, 1993; Potter & Faulconer, 1975) and in attending to and finding objects in visual search (Reeder & Peelen, 2013). How does attention deal with this complexity?

A number of studies indicate that simple features play a critical role for the selection of objects by showing that attention spreads throughout an

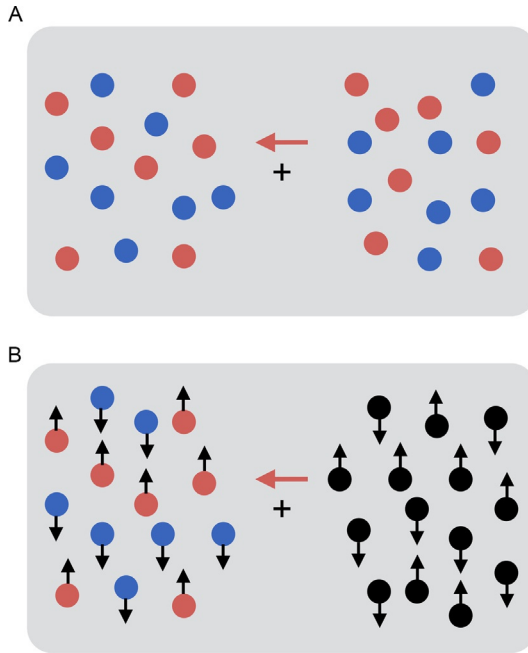

Fig. 1 Example displays from feature-based attention and multi-feature object studies, adapted from Chapman & Störmer, 2018 [preprint] (see also Störmer & Alvarez, 2014). (A) When participants are asked to attend to a color on one side (e.g., red on the left), visual processing of the attended color is enhanced throughout the visual field, even at task-irrelevant and unattended locations (e.g., the red dots on the right side of the display). (B) When participants are instructed to attend to one feature, for example, the color red, of a two-feature object or surface (all red dots are also moving upward), the secondary irrelevant feature (upward motion) is enhanced throughout the visual field, suggesting that attention to a single feature first spreads to other features of the same object but then can also spread across locations (Chapman & Störmer, 2018; preprint).

object, such that attending to a single visual feature of an object results in enhancement of all of that object's features, often in a seemingly obligatory fashion (Ernst et al., 2013; Katzner, Busse, & Treue, 2009; O'Craven, Downing, & Kanwisher, 1999; Schoenfeld et al., 2003). This within-object spreading of attention does not always seem to be confined by object boundaries, however. Studies from our lab and others have shown that features that are not directly attended but belong to an attended object can be globally enhanced (Arman, Ciaramitaro, & Boynton, 2006; Boehler, Schoenfeld, Heinze, & Hopf, 2011; Chapman & Störmer, 2018 [preprint]; Lustig & Beck, 2012; Melcher, Papathomas, & Vidnyánszky, 2005). For example,

when attending to the color of an object (e.g., red) that happens to move in a particular direction (e.g., upward), visual processing of that motion direction (upward) is enhanced throughout the visual field (Chapman & Störmer, 2018 [preprint]; Fig. 1B). Together, these studies indicate that feature-based attention can spread throughout objects, but that this spreading is not necessarily constrained to object boundaries, suggesting that simple visual features are crucial for the behavioral benefits of attention to more complex (multi-feature) objects.

While these studies investigate attentional selection for objects that consist of multiple features, moving one step closer to how attention operates in more complex settings, they often still use fairly simple visual objects (e.g., an object consisting of two simple features, e.g., color and motion). One benefit of this approach is that it is more straightforward to link the findings to the previous literature on feature-based attention. Yet, in our daily lives, selection is often based on more complex objects and object categories. For example, when looking for a person in a busy street scene, or when trying to detect an animal in the forest, we need to focus our attention on real-world object categories rather than one or two simple low-level features. How does attention operate across the vast object space that we are able to see and recognize?

One possibility is that when selecting complex objects, attention is tuned to the most diagnostic parts of that object and selection occurs based on these object parts or features rather than the entire object. For example, when looking for a car, attention would be tuned to the boxy shape and the roundness of the wheels (Evans & Treisman, 2005; Peelen & Kastner, 2014; Reeder & Peelen, 2013; Treisman, 2006), or when looking for fruits in the tree, attention would be tuned to their red color. But of course such part-based tuning can only occur when objects consist of diagnostic features, which is often not the case (target and distractors may share several lower-level visual features or shape parts). Thus, in many situations it would be most adaptive if attention was tuned to high-level object categories.

2.3 Attentional selection at the level of real-world objects and semantic categories

Several studies have investigated how attention to images of real-word objects, such as faces, houses, or cars, modulates visual processing and affects behavior. One standard paradigm consists of presenting two superimposed objects so that they compete at the same spatial location, and asking participants to attend to one of them (e.g., attend to a face overlaid on a

house image). The main result from these studies is that attention enhances neural responses in object-selective brain regions (Baldauf & Desimone, 2014; Cohen & Tong, 2015; O'Craven et al., 1999), demonstrating that attention can exert similar effects on the visual representations of complex objects as on those of simple features.

One core principle of the selection of simple visual features is that these features are enhanced across the visual field (see Section 2.2; Andersen et al., 2013; Sàenz et al., 2002; Serences & Boynton, 2007; Störmer & Alvarez, 2014; Treue & Martinez-Trujillo, 1999; Zhang & Luck, 2009). Does this global property also exist for real-world objects? In a recent study we investigated whether attention to complex real-world objects enhances object representations in a spatially global way (Störmer, Cohen, & Alvarez, in press). As a test case, we focused on the well-learned category of faces and examined a face-selective signal in the electroencephalogram (EEG) as a marker of category tuning (i.e., the face-selective N170). Across two experiments we found that face processing was enhanced when participants attended to faces (vs. buildings or houses) even at task-irrelevant and unattended locations (Fig. 2), suggesting that object-based attention can elicit similar, spatially-global effects on visual processing as has previously been found for simple features. To ensure that the observed effects were due to attending to the object category and not simple low-level features, we used stimuli that varied in low-level visual features (e.g., for faces: hairstyle, viewpoint, age, etc.; for buildings: houses, skyscrapers, towers; etc.) and included a control condition in which we asked participants to attend to scrambled face parts (instead of full-fledged faces). Critically, we only observed the spatial spreading of attention to other faces when participants were looking for complete faces but not when looking for face parts, indicating that attention enhances high-level visual processing across locations only when tuned to a specific feature configuration (the whole face) and not just to a few diagnostic parts of the object (e.g., the eyes). This is consistent with another functional magnetic resonance imaging (fMRI) study that showed spatially global effects of attention in object-selective cortex (e.g., lateral-occipital areas, LO) during a real-world visual search task in which participants were looking for people or cars (Peelen, Fei-Fei, & Kastner, 2009). While that study showed spatially global effects in brain regions sensitive to basic visual shapes (LO), our study extends these findings by demonstrating spatially global spreading of attention to even higher levels of visual processing, namely to the level of object categories (i.e., faces). Together, these studies suggest that attention to familiar and well-learned

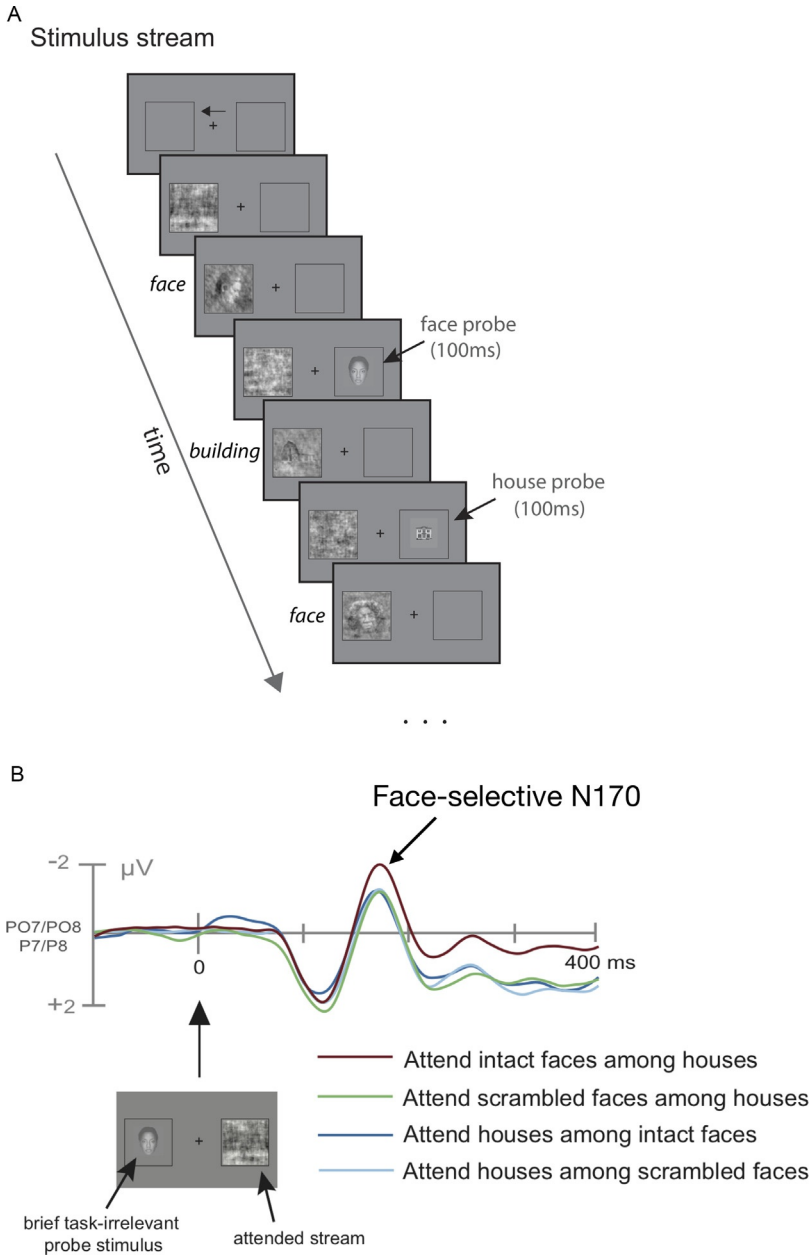

Fig. 2 (A) Task design and (B) data from Störmer et al. in press. When participants are asked to attend to faces at one location (e.g., left visual half-field) in a rapid stimulus stream, early visual processing of face probes presented on the unattended side (e.g., right visual half-field) is enhanced, as reflected by an amplitude increase of the face-sensitive N170 component in the EEG wave. This enhancement is only present when participants attend to full-fledged faces, and not when attending to scrambled face parts or buildings or houses.

object categories, like faces, people, and cars, can operate in a spatially global way similar to basic simple visual features.

Even though recent studies have begun to address the question of how attention operates across realistic objects, the categories and objects tested thus far have been restricted to a small set of stimuli, leaving open the question of whether similar effects hold for the vast number of objects that people can recognize. It is conceivable that attentional mechanisms depend on the representations over which they operate, which does not necessarily need to mean low- vs. high-level within the visual processing hierarchy, but could depend on the strength and robustness of the representations themselves. For example, if an object category is well-known, highly familiar, and often attended to, attention may operate on the full-fledged object (not just its parts or features) and result in similar effects as attention effects of low-level features (e.g., Störmer et al. in press). However, if a novel or less well-known object with a weaker underlying representation needs to be attended, attention may be tuned to the diagnostic parts of it (Evans & Treisman, 2005; Peelen & Kastner, 2014; Reeder & Peelen, 2013; Treisman, 2006).

Another line of research has used a different approach to address the question of how attention operates in more realistic viewing conditions. Instead of using a rather small but well-controlled stimulus set, one study used a very large (∼1000) set of object and action categories to investigate how attention modulates the representational semantic space across the whole brain (Cukur, Nishimoto, Huth, & Gallant, 2013). In that study, stimuli were presented in a movie clip while participants were looking for either people or vehicles. Using fMRI, the study measured how cortical responses across the brain changed depending on what category participants were searching for, and found that the brain representations were altered by attention. In particular, the amount of cortex activated by the attended category increased while neural activity corresponding to the task-irrelevant, unattended categories was reduced. These findings suggest that high-level attention can change the category selectivity of cortical voxels to be tuned toward the behaviorally relevant category at the expense of unattended irrelevant categories. Another neuroimaging study showed that attention increased the discriminability of neural population responses in higher-level visual areas. In that study, participants watched short videoclips of animals and either attended to their behavior (e.g., running, swimming) or to their taxonomy (e.g., insect, fish). It was found that the neural representations of the attended information were more discriminable relative to the unattended information in higher-level visual areas, as if attention

expanded the neural distances within the attended category of information (Nastase et al., 2017). Together, these studies are in agreement with some psychophysical data and theoretical accounts that have suggested that during visual search attention shifts tuning functions toward the attended target (Compte & Wang, 2005; Lee, Itti, Koch, & Braun, 1999; Lu & Dosher, 2004; Olshausen, Anderson, & Van Essen, 1993).

While these fMRI studies demonstrate that attention to object categories can have large-scale effects across the brain, it remains an open question what the underlying neural mechanisms are. One possibility is that these modulations are due to selective increases in neural gain operating differentially on subpopulations of neurons contained within a voxel. Another possibility is that they are the direct consequence of shifts in categorical tuning at the neural level. Both mechanisms have been shown to exist for simple, low-level features (David et al., 2008; Martinez-Trujillo & Treue, 2004; McAdams & Maunsell, 2000; Motter, 1994), but it yet needs to be determined what happens at the level of object categories and in higher-level brain regions.

Thus, taken together, a large amount of work on visual attention has focused on attention to simple features. While understanding how this scales up to fully realistic scenes is quite challenging, newer work has suggested that attention can operate not only over basic features like color but also over categories of objects, and appears to show similar tuning properties in some cases, for example, the spatial spreading of attention to faces. Furthermore, recent findings suggest that feature-attention in general not only enhances relevant features of the attended objects but also spreads to other features of attended objects and across locations, enhancing these features. However, there remain many questions about the nature of visual attention, feature attention and object-based attention as applied to realistic objects and scenes.

2.4 Attention to, and processing of, "ensembles" and groups of objects

In addition to representing individual features or objects, people can quickly and accurately compute summary statistics of simple visual features, like the mean size of the items (Ariely, 2001; Chong & Treisman, 2003), the mean orientation (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001), and the mean location (Alvarez & Oliva, 2008). People are also capable of representing more than just the mean, for example, representations of variance (Solomon, Morgan, & Chubb, 2011). In these types of tasks people are usually briefly presented with an array of items (e.g., oriented lines, different sized stimuli) and are asked to subsequently report the average

orientation or size of the set of stimuli they just saw. Importantly, this kind of ensemble processing goes beyond the typical basic attentional paradigm by providing evidence that despite the best intention of the experimenters, people often will not treat items independently but also process information about the "ensemble" of the display even in displays of simple stimuli (e.g., Brady & Alvarez, 2011). However, the ability of people to encode information about entire sets of objects provides the opportunity to examine more real-world use cases of attention (e.g., dividing attention between individuals or processing items as an entire set) while maintaining control of stimuli.

Several studies have also shown that ensemble perception can operate on sets of high-level objects, rather than simple visual features, with a particular focus on faces. That is, observers can report the average emotion and gender of a briefly presented group of faces (Haberman & Whitney, 2007, 2009) as well as the average facial identity of a group (de Fockert & Wolfenstein, 2009; Neumann, Schweinberger, & Burton, 2013; Yamanashi Leib et al., 2012), and facial attractiveness (Walker & Vul, 2014). To evaluate whether these effects occur at a high level of processing (i.e., holistic face processing) and are not driven by averaging low-level visual features (e.g., the whiteness of the teeth, the roundness of the eyes, etc.), some studies have tested whether ensemble statistics can also be extracted from inverted or scrambled faces, and have generally found that this is not as accurate (Haberman & Whitney, 2009; Yamanashi Leib et al., 2012). Consistent with the interpretation that these summary statistics can be formed at a configural and holistic face processing stage, recent studies have also shown that these facial summary statistics can be computed over viewpoint-invariant representations and not just 2-D image level information (Neumann et al., 2013; Yamanashi Leib et al., 2014). In other words, people can compute summary statistics like "average identity" even across sets of faces that differ in viewpoint from each other (and thus differ significantly in low-level features).

In addition to facial expressions, gender, and identity, it has also been shown that participants are able to report the average gaze direction and mean head rotation of crowds of people (Florey, Clifford, Dakin, & Mareschal, 2016; Sweeny & Whitney, 2014), as well as the mean biological motion direction of point-light walkers (Sweeny, Haroz, & Whitney, 2013), although all with considerably smaller set sizes (e.g., processing only a few items, rather than many at once). Based on these findings, it has been suggested that ensemble perception may be particularly useful for high-level visual perception to guide social interactions, for example, by facilitating the

recognition of crowd panic that may involve the average and variance of heading direction, motion speed, and average emotion (Whitney & Yamanashi Leib, 2018).

It remains an open question how people compute ensemble statistics, how they are related to attention, and the extent to which they are directly perceptually available. For example, it may be that whenever you grab many objects with your attention, you automatically get an ensemble representation of those objects. Some have even proposed a more dramatic departure from previous theories, with summary statistic representations being the core cause of "attentional" capacity limits. It remains unclear exactly how ensemble representations relate to attention. It also remains unclear whether the high-level "ensemble" effects (e.g., faces) reflect the same type of summary representation and involve similar computations as ensembles for sets of simple stimuli (e.g., motion directions, orientations).

There is some reason to believe that at least some ensemble tasks do not tap into a fundamentally different representation than tasks where people are asked about just a single item. For example, in our own work, we have shown that much of the noise in ensemble computation comes from perceptual noise—e.g., people with noisy representations of a single gabor orientation also have noisy representations of the mean orientation of a set of gabor stimuli (Haberman, Brady, & Alvarez, 2015). Furthermore, recent work from our labs has shown that under some circumstances, people may use attention strategically to answer questions about ensemble properties (like the variance in size of a set of objects), in effect converting the question into one they can answer using visual search (Lau & Brady, 2018a, 2018b). For example, we show that participants seem to rely heavily on the range of the set of objects (the largest and smallest object) rather than directly computing variance, and seem to perform visual search to find these objects, leading to more reliance on items near fixation and less reliance on objects far from fixation. Thus, we believe that much work remains to be done to address exactly what strategies and representations participants use when they are asked to report the mean or variance of a set. Their ability to report something approximately like a mean does not imply they necessarily have a direct representation of the mean; in many cases they may be relying upon other, seemingly complex strategies that ultimately give something that approximates a statistical summary representation (e.g., Myczek & Simons, 2008).

Interestingly, while the majority of the work on "ensemble" representations (e.g., simultaneous representation of information about the entire set or ensemble of items) has focused on statistical summaries like the mean,

ensemble representations do not always collapse information to a single point estimate (e.g., a mean or variance) but instead sometimes preserve spatial information. When considering more real-world scenarios—where there is not just a single set of relatively homogeneous stimuli on a white background—it almost seems necessary that some spatial information would necessarily be preserved in computing ensemble information. This class of spatial ensemble representations includes the spatial distribution of orientations (Alvarez & Oliva, 2009); the relative homogeneity of a display (Victor & Conte, 2004); and basic texture statistics (Brady & Tenenbaum, 2010, 2013). These spatial ensemble representations are quickly and accurately computed, even outside the focus of spatial attention. As an example of such a representation, Fig. 3 shows a sample spatial ensemble task. In such a task, participants are better at detecting changes that flip the ensemble structure (e.g., flip the vertical/horizontal arrangement) then those that do not, suggesting that they represent more than just individual objects, and more than just average orientation; they also encode a spatial ensemble representation (Alvarez & Oliva, 2009; Brady et al., 2017).

Why are participants able to compute spatial ensemble representations so quickly, even without focal attention (Alvarez & Oliva, 2009) and in the visual periphery (Balas, Nakano, & Rosenholtz, 2009; Rosenholtz, Huang, Raj, Balas, & Ilie, 2012)? One idea has been that these properties are consistent with a potential role of such representations in rapid scene recognition. The spatial ensemble patterns people are most sensitive to appear to closely mimic the patterns of oriented elements traditionally used in computer vision algorithms to holistically recognize scenes (e.g., Oliva & Torralba, 2006). It is thus possible that human sensitivity to these patterns in simple displays, like grids of gabors and displays of colored circles, arises because such patterns mimic the mid-level features people use to recognize surfaces and the spatial layout of scenes (Brady et al., 2017).

Together, we believe work on ensemble perception—both of summary statistics and spatial ensembles—provides an important bridge between simple stimuli and real-world scenes. The kinds of distributed patterns of information across objects that are typically studied in ensemble tasks are likely a major part of real-world recognition and how we use and distribute our visual attention in the world.

2.5 Cross-modal influences on perception and attention: Real-world objects make sounds

In the real world, visual objects are often accompanied by sounds, smells, tactile information, or taste. How does the human brain integrate the inputs

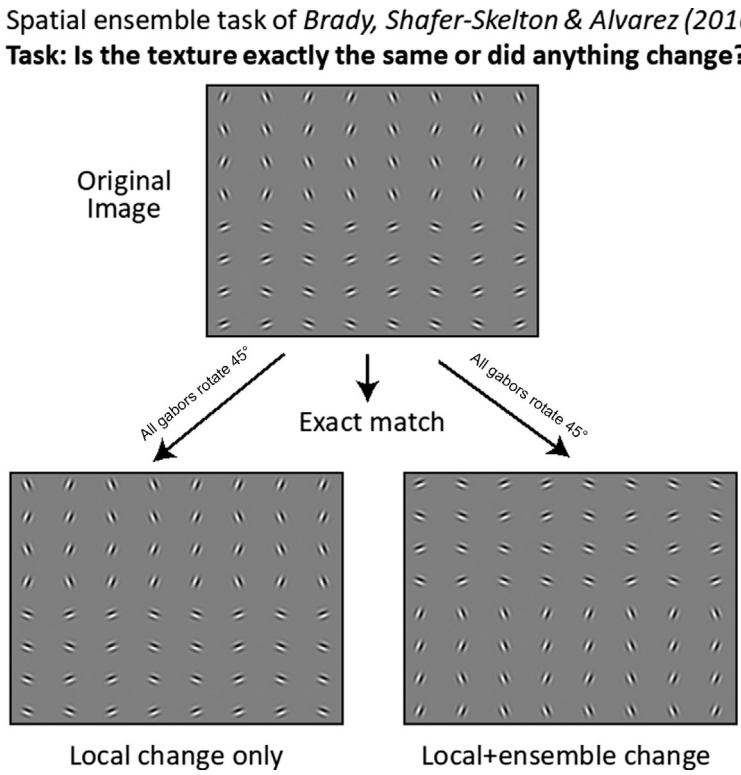

Fig. 3 A sample spatial ensemble task from Brady, Shafer-Skelton, and Alvarez (2017), adapted from Alvarez and Oliva (2009). In this task, participants are briefly shown a grid of 8×8 gabors (for <100 ms) while their attention is spread diffusely. This display is unique in that there is a large-scale structure in the gabor elements: the top ones are vertical-ish and the bottom ones horizontal-ish. Note that this is a pattern of high spatial frequency—there is no information in the low spatial frequencies; blurring the display results in a solid gray field. After seeing the stimulus, there is then a delay interval, followed by the reappearance of the gabors. Sometimes the display is exactly the same; but sometimes every gabor element has rotated by 45°. Crucially, this can result in either the same ensemble structure (local change only) or can flip the spatial ensemble structure (local + ensemble change). Participants are considerably better at detecting changes that flip the ensemble structure, suggesting that they represent more than just individual objects, and more than just average orientation; they also encode a spatial ensemble representation. In addition, participants who are most benefited by the ensemble structure are also the best at rapidly recognizing visual scenes (Brady et al., 2017).

across these different sensory modalities to form coherent, multisensory representations, and what role does attention play? Previous research has often focused on understanding perception and attention through a single, isolated sensory modality. While this research has provided us with

substantial knowledge on basic mechanisms of attention, it misses important aspects of the real world, where the sound of a car beeping may be critical not just to orienting our attention to an important location in the world, but may even prime us to be ready to recognize a car when we direct our attention to that location. Therefore, to better understand visual perception and visual attention as a whole it is vital that we understand how perception and attention integrate information across different sensory modalities.

Some of the seminal work on multisensory integration has shown that highly specialized neurons alter their firing patterns in superadditive, additive, or subadditive ways when in the presence of cross-modal audiovisual input. A pioneering study by Meredith and Stein (1983) that used simple, low-level auditory (a short hissing sound) and visual (a small bar of light) stimuli showed that neurons in the cat superior colliculus (a midbrain region) respond with roughly equal firing rates to both stimuli. Interestingly, however, when these stimuli were presented cross-modally (i.e., simultaneously), some neurons produced an integrated response that substantially exceeded the sum of their responses to unimodal stimuli while others showed greatly reduced responses, indicating that they are specific multisensory neurons that are sensitive to cross-modal inputs. Building on this work, multisensory neurons that respond to more complex, naturalistic stimuli have been identified in various brain regions across nonhuman species (Ghazanfar, Maier, Hoffman, & Logothetis, 2005; Wallace, Ramachandran, & Stein, 2004; for a review see Stein & Stanford, 2008). For example, one study separated the auditory and visual components from videos of rhesus monkey vocalizations and showed that when the stimuli were presented cross-modally, neurons in the rhesus ventrolateral prefrontal cortex responded with firing rates that were far greater than responses to the unimodal stimuli alone (Sugihara, Diltz, Averbeck, & Romanski, 2006). Thus, there is significant evidence showing that particular neural populations are sensitive to cross-modal inputs.

Other studies have shown that multisensory influences can also occur in early sensory areas that are traditionally thought of as being unisensory. For example, we recently showed that hearing a salient sound can activate human visual cortex, even when the sounds are task-irrelevant (McDonald, Störmer, Martinez, Feng, & Hillyard, 2013; Störmer, Feng, Martinez, McDonald, & Hillyard, 2016). Interestingly, these cross-modal effects occurred in a spatially specific way such that a sound from the left side of space elicited activity in right visual cortex, and vice versa. The spatially selective nature of these cross-modal effects suggests that spatial

processing—or more specifically, spatial orienting of attention—is inherently linked to the visual cortex. What are the behavioral consequences of these sound-induced changes in visual excitability? A series of studies have shown that sounds can facilitate visual processing in different ways: peripheral sounds increase the detection of faint visual stimuli at the location of the sound (Dufour, 1999; Frassinetti, Bolognini, & Làdavas, 2002; McDonald, Teder-Sälejärvi, & Hillyard, 2000); enhance contrast perception (Störmer, McDonald, & Hillyard, 2009); facilitate visual discrimination (Feng, Störmer, Martinez, McDonald, & Hillyard, 2014); and accelerate the perceived timing of visual stimuli (Mcdonald, Teder-Sälejärvi, Russo, & Hillyard, 2003; Zampini, Shore, & Spence, 2005). Thus, orienting attention to a salient peripheral sound can have multiple influences on visual perception, suggesting that spatial attention operates across auditory and visual sensory modalities. Given that in the real world visual objects are often preceded by sounds (e.g., the honk of a car alerts us to where a car is about to appear in our field of view), it seems particularly adaptive that attention to auditory inputs is rapidly transmitted to visual areas.

Other studies have shown that sounds also influence visual perception in a spatially non-specific way. For example, when a single light flash is presented centrally with two tones, people often report seeing two light flashes (Mishra, Martinez, Sejnowski, & Hillyard, 2007; Shams, Kamitani, & Shimojo, 2000). Furthermore, visual target detection has been shown to be improved when an uninformative tone burst is presented during a difficult visual search task (Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008). Giard and Peronnet (1999) showed that after having learned audio-visual pairings, simple object recognition (e.g., vertically vs. horizontally oriented oval) was faster and more accurate in the presence of the paired audio-visual stimuli compared to unimodal trials. Using EEG they also demonstrated that the integration of these signals happens at an early stage of sensory processing. Together, these studies demonstrate that visual processing is strongly influenced by inputs from other sensory modalities (in particular, audition, as reviewed here); thus, to fully understand visual perception and attention, it is important to consider these multisensory influences.

Many studies reviewed so far focus on multisensory influences for simple, low-level stimuli such as noise or tone bursts and basic shapes. How is sensory information from multiple modalities combined for complex, real-world stimuli such as the sound of a bird tweet and the sight of bird? One possibility is that real-world sounds are highly specific to a particular object and so are less able to facilitate visual processing for object

categories than other, more abstract cues (Edmiston & Lupyan, 2015; Lupyan & Thompson-Schill, 2012). In line with this, Boutonnet and Lupyan (2015) recorded EEG while participants rapidly categorized an image as being either congruent or incongruent with a previously heard auditory stimulus and showed that participants were slower, and showed less robust early sensory processing, when the auditory cue was a real-world sound (e.g., a dog's bark) compared to a spoken-word (e.g., "dog"). However, it is possible that real-world sounds nevertheless influence multiple stages of visual processing thereby activating higher-level, semantic-level features (Chen & Spence, 2010; Iordanescu, Guzman-Martinez, Grabowecky, & Suzuki, 2008; Schneider, Engel, & Debener, 2008), even if they activate more specific features, on average, than do words (Boutonnet & Lupyan, 2015). Consistent with this account, a recent fMRI study showed that real-world auditory stimuli can activate early visual cortex reliably enough to decode and accurately predict which auditory stimuli are being heard (Vetter, Smith, & Muckli, 2014). Similar to what has been shown for simple stimuli (e.g., McDonald et al., 2013), it appears that the timing between auditory and visual inputs is critical to elicit any multisensory effects. For example, Chen and Spence (2011) demonstrated that the degree to which real-world compared to spoken-word auditory stimuli exert influence on high-level visual processing was highly dependent on when (early vs. late) the auditory stimuli were presented in relation to the visual objects.

In sum, recent work points to critical influences of auditory stimuli on visual processing, both for simple and for complex, real-world objects, highlighting the importance of trying to understand perceptual and attentional processes across sensory modalities. While many studies point to important links between these modalities, it is still unclear what the mechanisms are by which information is integrated across the senses, and whether similar mechanisms play a role for simple, low-level stimuli as for complex, real-world objects. In future work, it will be important to expand research on visual perception and visual attention to also consider multisensory influences that help structure neural dynamics and behavior in real-world settings. Broadly, scaling up to incorporate crossmodal integration will be critical to understanding visual attention in the real-world.

2.6 Visual attention conclusion

In this section we have argued that visual attention can be usefully studied by scaling up from the simplest settings to more complex settings that capture

some aspects of the complexity of the real-world, while still remaining in the realm of well-controlled stimuli and well-understood tasks. We focused on only a limited subset of attention tasks, particularly those that are not traditionally tightly coupled to real world problems, but for which there is the potential to scale-up to more realistic scenarios. Indeed, many realistic attention tasks have been well studied throughout the literature in visual attention, which we do not focus on in this chapter. For example, multiple object tracking (Pylyshyn & Storm, 1988) is a common everyday task, and has been well studied, including the extent to which performance and abilities at this task vary across populations (Störmer, Li, Heekeren, & Lindenberger, 2013) and are affected by intuitive physical knowledge about how objects move in the world (Lau & Brady, 2018b [preprint]). Similarly, visual search is an extremely common real-world attention task, and is well studied in the literature on visual attention (e.g., Wolfe & Horowitz, 2004), including relevant work on the attentional mechanisms that underlie search (e.g., which items are hardest to search among: Störmer & Alvarez, 2014) as well as how we learn across searches (e.g., Brady & Chun, 2007; Chun & Jiang, 1998). Thus, there is a rich history of well-controlled visual attention literature aimed at more real-world tasks. We have argued that the literature on basic mechanisms and limits of attention can also benefit from a consideration of more real-world problems (e.g., crossmodal integration; summary statistics across many objects; etc.).

3. Visual working memory

3.1 Introduction

Visual working memory is a system used to actively store and manipulate visual information (Baddeley, 2012). This memory system is severely limited in capacity (Cowan, 2001), and the capacity limits of the active storage system in particular are closely related to measures of intelligence and academic achievement (Alloway & Alloway, 2010; Cowan, 2005; Fukuda, Vogel, Mayr, & Awh, 2010), suggesting that active storage in visual working memory may be a core cognitive ability that underlies, and constrains, our ability to process information across domains (Brady, Konkle, & Alvarez, 2011).

The vast majority of studies on visual working memory have focused on memory for simple stimuli like colored squares, oriented lines or novel shapes; these are all stimuli about which participants have minimal background knowledge or expectations. These simple, meaningless stimuli are

assumed to best assess the core capacity of working memory because they have no semantic associations and are repeated from trial-to-trial, which minimizes participants' ability to use other memory systems, like episodic visual long-term memory (Cowan, 2001; Lin & Luck, 2012). Episodic long-term memory is the process of forming memory traces and later retrieving them without continued active maintenance, and it can be used at any time scale (even with brief delays). Contributions from this system, which operates best with conceptually meaningful stimuli and when there is little interference from items repeating across trials (Konkle, Brady, Alvarez, & Oliva, 2010; Wickens, Born, & Allen, 1963; Wiseman & Neisser, 1974), are thought to be minimal in working memory tasks that use simple, meaningless stimuli.

However, while studies using such simplified stimuli have provided critical insights into the structure of the working memory system and the nature of its capacity, they also leave out many important aspects of visual memory in the real world. In particular, in the remaining part of the chapter we argue that (1) meaningfulness, knowledge, and familiarity play an important role in visual working memory and in shaping its capacity; (2) memory for scenes and surfaces is critical in the world, not just memory for objects; (3) even in the case of memory for objects, important regularities between objects often give rise to ensemble information, rather than just individual object representations. Many of these factors are rarely studied in the context of visual working memory, and when they are it is often with stimuli that—while having extremely strong external validity—are difficult to fully understand in terms of representations and processes (e.g., memory for totally real-world scenes; Hollingworth, 2004). How does visual working memory function when these more realistic factors are present? Can we understand this in well-controlled experiments where we can build computational models and understand the memory representations participants form?

3.2 Visual working memory stores not just individual items, but ensemble information

The vast majority of visual working memory studies attempt to isolate individual object representations and understand capacity in terms of individual objects. In a typical visual working memory display, participants see several simple, isolated objects on a solid background and are asked to hold these items briefly in mind, and then either detect whether any of the objects changed when they reappear (Luck & Vogel, 1997) or report one of the items from memory (Wilken & Ma, 2004; Zhang & Luck, 2008).

Experiments like this reveal a stark capacity limit: participants are able to hold in mind only a few objects under most conditions. Over the past 20 years, a huge amount of research has used paradigms like this one to investigate the most important issues that arise when considering how people remember individual objects. For example, many researchers have focused on how flexibly we can allocate our working memory resources to different numbers of objects (e.g., "slots" vs. "resources"; Alvarez & Cavanagh, 2004; Bays, Catalao, & Husain, 2009; Schurgin, Wixted, & Brady, 2018 [preprint]; Zhang & Luck, 2008). Another major area of work has demonstrated that visual working memory capacity, even in simple displays, is predictive of fluid intelligence as well as other important cognitive abilities (Fukuda et al., 2010; Unsworth, Fukuda, Awh, & Vogel, 2014).

Surprisingly little work has examined the relationship between ensemble representations—discussed in Section 2.4—and the individual items that are stored in visual working memory. The existing work that has examined this connection has often found that ensemble structure plays an important role in the representation of even simple visual working memory displays (e.g., Brady & Alvarez, 2011, 2015b; Brady & Tenenbaum, 2013; Orhan & Jacobs, 2013; Sims, Jacobs, & Knill, 2012; Swan & Wyble, 2014). Thus, although typical visual working memory displays are, as best as possible, prevented from having any overarching structure or gist, nevertheless participants seem to make use of not only individual object representations but also ensemble representations when encoding these displays. This means that items in visual working memory are not encoded independently (Brady et al., 2011; see also Jiang, Olson, & Chun, 2000; Johnson, Spencer, Luck, & Schoner, 2009; Lin & Luck, 2009). This is true both in the case of explicit perceptual grouping, where items are combined and treated as though they were a single perceptual unit (e.g., Morey, Cong, Zheng, Price, & Morey, 2015), and in cases where it appears participants do encode individual items separately but also encode ensemble information (e.g., Brady & Alvarez, 2011, 2015a, 2015b).

Understanding the relationship between ensemble representations and individual objects is important for the insight it provides into visual working memory representations, and also important because the use of ensemble structure has frequently been overlooked in existing visual working memory experiments, leading to vastly different characterizations of the cognitive architecture. This is true for both simple ensemble representations—leading to memories that integrate item and ensemble information even when not

required (Brady & Alvarez, 2011)—and spatial ensemble representations (e.g., those that preserve spatial structure, akin to peripheral texture representations).

One of the most well-cited papers in the visual working memory literature provides an example of a spatial ensemble effect (Awh et al., 2007). In this paper, the authors argue for a "slot" model of visual working memory capacity. They suggest that people always represent three to four objects and argue that only the precision with which these objects are represented is affected by how complex these objects are. Supporting this, they find that people can easily detect large changes to complex objects (e.g., a cube changing to a Chinese character; see Fig. 4), with performance consistent with remembering four individual objects. But small changes (changes from a cube to another cube) often cannot be detected. This is taken as evidence

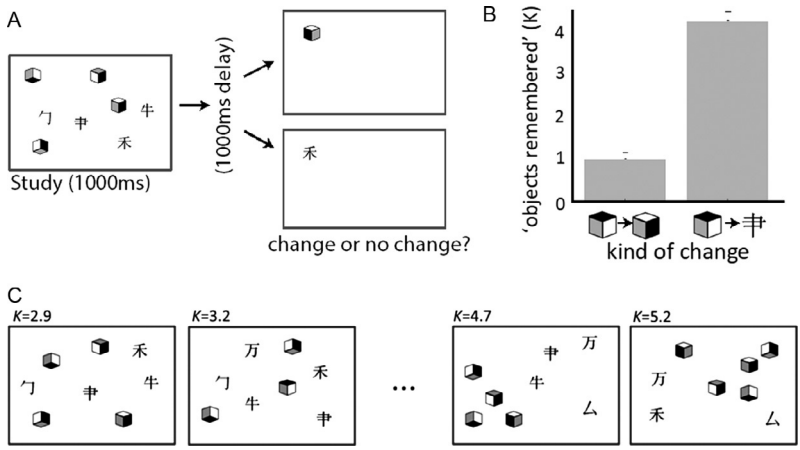

Fig. 4 (A) Participants performed a change detection task with either within-category (small) or cross-category (large) changes. All participants saw exactly the same initial displays but were tested on different items. (B) Replicating Awh, Barton, and Vogel (2007), participants seemed to remember only one object well enough to detect within-category changes, but four objects with sufficient fidelity to detect large changes. Error bars correspond to 1 standard error of the mean. (C) The individual displays where participants performed best on the cross-category changes (right) were those in which the cubes were clustered together, such that participants could detect a change from a cube to a character based on a change in clustering (an ensemble or texture representation) rather than an individual item memory. The figure shows the two worst displays (left) and two best displays (right) for illustration of this effect, with the capacity estimate for cross-category changes (K) for each display listed above it. Note that even the worst displays still contain fairly significant ensemble information, since they have only two kinds of objects present.

of low-resolution representations, and a capacity limit of only ~1 high-resolution object. However, one key assumption of this work is that performance is supported solely by the maintenance of individual objects, with no contribution from ensembles. In recent work (Brady & Alvarez, 2015a), we have shown that this is incorrect, and that the claim that participants remember four individual, low-resolution items does not capture the representations formed of these displays. For example, an item analysis (see Fig. 4), which looks at which displays were easiest and hardest for participants to detect large changes in (cube → character), shows that the displays where participants perform best are actually the displays where participants are most likely to rely on ensemble representations. Participants rely on representations of the spatial structure of these displays (e.g., on the far right display in Fig. 4C, the top right of the display is darker and "heavier" than the remainder of the display). Once ensembles are taken into account, it is clear that fewer complex objects are individually stored than expected. This is in contrast to strong claims of fixed capacity required by slot models, and thus provides evidence for a more continuous-resource-like architecture for visual working memory. These results show the importance of taking into account ensemble representations when characterizing working memory, rather than treating all information participants remember as though it arises solely from individual object representations.

In further work (Schurgin & Brady, 2019) we have shown that it is also the case that this use of spatial ensemble representations results in a non-fixed capacity estimate, again arguing against slot-like views. That is, people seemingly remember more information with more items present, contrary to a fixed capacity view of working memory; however, this appears to be because ensemble information becomes more important at larger set sizes. In particular, changes to the stimuli that do not affect the similarity of individual items, but do affect how well ensemble representations capture the difference between stimuli, result in significant changes in calculated memory "capacity" (Schurgin & Brady, 2019). Thus, we believe it is of fundamental importance for models of visual working memory to consider that in nearly every display with multiple objects, people do not encode items independently, and thus we cannot conceive of the capacity of working memory without considering how people use ensemble and other relational information, and how this information about multiple items integrates with information about single items (e.g., Chunharas, Rademaker, Brady, & Serences, 2019 [preprint]).

Broadly, then, a growing body of work shows that people do not remember items in working memory independently. Perceptual grouping, ensemble statistics/summary statistics, spatial ensemble representations, and relational encoding between items may all play a significant role in memory even in the simplest displays. Understanding the way people use these representations and possibly rely on them in strategic ways—and how this may scale up to real-world scenes—thus remains an important area of work for visual working memory researchers.

3.3 The world has more than objects: Scenes and surfaces in visual memory

While most working memory literature focuses on studying memory for discrete objects or sometimes ensembles of multiple objects, both neuroimaging and behavioral studies find evidence for a general dissociation between perceptual processing of discrete objects and the larger surfaces that make up a scene. For example, a large amount of work has shown that there are scene-selective brain regions in humans, which respond selectively to scenes compared to objects (Epstein, 2005; Epstein & Kanwisher, 1998; Kravitz, Saleem, Baker, & Mishkin, 2011) and which seem to represent features of a scene's spatial layout (e.g., openness) rather than the objects it contains (Epstein, 2005; Lescroart & Gallant, 2019; Park, Brady, Greene, & Oliva, 2011). In addition, behavioral experiments show that it is possible to recognize briefly presented scenes even without being able to recognize any of the objects in those scenes (Oliva & Torralba, 2001; Schyns & Oliva, 1994), providing evidence of the independence of scene recognition from object recognition. Does this distinction hold in working memory? The sensory recruitment view of working memory suggests that perceptual regions are involved in working memory (e.g., Serences, 2016; Serences, Ester, Vogel, & Awh, 2009), so the perceptual distinction between objects and scene representations may well persist into working memory.

While very little work successfully dissociates effects of scene layout from those of objects or low-level features like orientation, some studies have made progress in understanding memory in natural scenes. For example, it has been shown that previews of scenes facilitate subsequent processing related to that scene, including subsequent visual search for an object present in that scene (Castelhano & Henderson, 2007; Võ & Henderson, 2010). The memory representations retained in these studies are not totally low-level; that is, they are abstracted from the exact visual features (e.g., Castelhano & Henderson, 2007 show size invariance). However, these

studies do not make it clear what specifically about the scene is remembered across the delay or to what extent this memory reflects the layout of the major surfaces of the scene as opposed to hypotheses about particular objects and their locations.

A different literature has aimed to test memory for scene layout, but without a focus on visual working memory per se (as it has used very short delays). In particular, Sanocki and colleagues have examined memory for—and preview benefits related to—the 3D spatial layout of a scene. They have done this using a paradigm where a preview of the scene is shown (with only some information present, largely layout information), and asked whether such previews facilitate a depth judgment on two items that subsequently appear within that scene (e.g., Sanocki, 2003, 2013; Sanocki & Epstein, 1997; Sanocki, Michelet, Sellers, & Reynolds, 2006). The idea is that participants' task—deciding which of two objects in a scene is closer in depth to the viewer—specifically targets scene layout representation, requiring participants to have remembered which parts of a scene are near or far from the observer (as opposed to only having held in mind a distribution of possible locations of objects). This "scene priming" paradigm is widely cited as an example of scene layout information being maintained in memory (e.g., by Chun & Jiang, 1998; Oliva & Torralba, 2001).

However, the effect is often diminished when some low-level information is varied (Sanocki, 2003), and recent work from our labs calls into question the original interpretation of these studies as reflecting memory for scene layout (Shafer-Skelton & Brady, 2019). Most notably, the effect disappears when a longer (200 ms) delay and a mask separate the preview and target images, preventing high-capacity fragile memory such as iconic memory from contributing to the effect (e.g., by making the target dots easier to find). Further experiments suggest that these effects may result from participants being faster to detect the target dots after a relevant preview (due to shared features between the preview and target images facilitating the pop-out of the target dots), rather than being faster at performing the depth judgment (Shafer-Skelton & Brady, 2019). These results dramatically broaden the space of possible interpretations of these paradigms, calling for new behavioral work establishing whether scene layout information can be maintained in working memory.

One critical motivation for this work is to better understand the role that memory for scenes and surfaces could play in constructing a (seemingly) complete view of the world from the smaller views we perceive with each fixation. As we move our eyes, it is often claimed that nearly all information

about the world is lost between subsequent fixations, with only information already in working memory or long-term memory persisting to form an integrated scene representation (Aagten-Murphy & Bays, 2018; Luck, 2008). Yet working memory's limited capacity makes this seem too impoverished to result in the robust scene representations we ultimately form. One idea is that in addition to objects being stored in working memory, as nearly always studied in the literature, we also store scene information like spatial layout. Because of evidence that some scene information may be extracted more quickly than object information (Greene & Oliva, 2009) and processed with less attention (Li, VanRullen, Koch, & Perona, 2002), scene-specific information may provide a much stronger basis for constructing a complete view of the world across saccades.

Understanding working memory for scene layout is also important for studying working memory capacity in general. While a great deal of work has focused on quantifying memory for simple shapes in the same domain, they may underestimate our capacity in the real world, where there are many diverse types of information that could draw on distinct resources for their maintenance.

While much remains to be done in understanding the role of surfaces in spatial layout and the way these are stored in memory, some work from our lab has looked at the benefit of depth itself on visual working memory. For example, in Chunharas, Rademaker, Sprague, Brady, and Serences (2019), we recently showed that separating memoranda in depth increases visual working memory performance. That is, when items are on different depth planes, they appear to interfere with each other less. This suggests a role for depth—and possible surface segmentation more broadly—in shaping what is remembered in visual working memory.

3.4 We have knowledge about the objects we wish to remember: Learning and knowledge in visual working memory

Working memory training has become an important topic, in particular the possibility that training on working memory tasks will transfer not only to different stimuli but also increase other abilities like fluid intelligence (Jaeggi, Buschkuehl, Jonides, & Perrig, 2008). The ability of working memory training to transfer to other abilities is controversial (Chooi & Thompson, 2012; Shipstead, Redick, & Engle, 2012; Thompson et al., 2013). However, most studies find that persistent training on a particular working memory task improves performance on that task (e.g., Thompson et al., 2013).

In visual working memory, studies have investigated stimulus effects, specifically the effects of repeated exposure to the same memory displays on visual working memory capacity (Olson & Jiang, 2004; Olson, Jiang, & Moore, 2005). However, they have not provided clear evidence that familiarity with the stimuli can increase capacity. For example, one study found evidence that learning did not increase the amount of information remembered, but that it improved memory performance by redirecting attention to the items that were subsequently tested (Olson et al., 2005).

However, we have shown that visual working memory performance is reliably improved when participants are taught novel associations between otherwise-meaningless elements (e.g., colored circles). For example, Brady, Konkle, and Alvarez (2009) found that repeated exposure to pairs of stimuli (e.g., red next to yellow) led participants to remember more colors as they learned this association. This improved capacity was evident even after considering the possibility that such associations could allow for perceptual "guessing" strategies. For example, Brady, Konkle, and Alvarez (2009) demonstrated that when an item happened to not be in its usual pair (red happened to be next to blue, not yellow), participants did not falsely guess "yellow" for the item next to red. These capacity improvements occurred even though none of the elements were connected to broader conceptual meaning and thus provide some evidence that exposure or familiarity alone could improve working memory capacity (in the context of paired stimuli). Thus, at least with simple stimuli like letters and colors, while exposure to independent stimuli alone does not increase performance in working memory (Chen, Yee Eng, & Jiang, 2006; Eng, Chen, & Jiang, 2005; Olson & Jiang, 2004), building items into larger associative units can improve working memory.

What about the case of memory for realistic objects? In the real world, we almost always remember objects that don't only consist of associated pairs of visual features (as in Brady, Konkle, & Alvarez, 2009), but also these objects themselves are meaningful and connect to existing semantic knowledge. How does such knowledge impact visual working memory?

In experiments using simple stimuli, the active working memory system is often estimated to have a fixed capacity no matter how long participants are given to encode those items (Luck & Vogel, 1997), a capacity sometimes deemed to depend on the limits of attentional processes, in line with its active nature (Cowan, 2001). However, by contrast to the fixed capacity that is observed with simple stimuli, similar tasks with real-world objects have found that participants remember more items with more time, without

an obvious capacity limit (Brady, Konkle, Oliva, & Alvarez, 2009; Melcher, 2001, 2006). Furthermore, paradigms with real-world objects where interference is minimized also show that participants can remember a very large number of objects without reaching a fixed capacity limit (Endress & Potter, 2014).

These differences in capacity estimates for simple stimuli and real-world objects could simply be due to the fact that working memory operates equally well on both stimulus types, but real-world objects can additionally benefit from the high-capacity episodic long-term memory system or a form of more accessible long-term memories (Cowan, 1988). For example, many theories suggest that with expertise and familiarity, non-active forms of memory may become quickly available and thus may be used for performing tasks (e.g., Ericsson & Kintsch, 1995). This suggests that performance could be enhanced with expertise and familiarity by a working memory system that may be non-active and does not depend on sustained neural firing. Alternatively, it is possible that, at least to some extent, the active working memory system has a different (and higher) capacity for real-world stimuli than for simple stimuli.

Recent work from our labs has made use of a neural correlate of visual working memory—observable in the EEG—called the contralateral delay activity (CDA), to examine whether or not the maintenance of real-world objects relies on the same active working memory system as simple stimuli. Behavioral data have consistently revealed that with long encoding times, participants are able to remember more real-world objects than simple colors, despite the fact that the real-world objects are more complex (Brady, Konkle, Oliva, et al., 2009; Brady, Störmer, & Alvarez, 2016; Melcher, 2006). Is this additional information a result of storage in visual working memory systems, or is it a result of performance being enhanced by the use of non-active memory systems like episodic long-term memory? To examine this, we directly compared the CDA when people were remembering colors and real-world objects (Brady et al., 2016). We found that the CDA was reliably greater for remembering five objects than for remembering five colors (but not different when the amount remembered was the same for each stimulus set, e.g., with three of each presented). Because the CDA indexes working memory storage in particular, this suggests that the behavioral finding that additional real-world objects are remembered beyond the limit on color memory is at least partially the result of active storage in visual working memory systems, rather than being solely due to the use of the episodic long-term memory system (Brady et al., 2016)

or other forms of non-active storage (Ericsson & Kintsch, 1995). However, it remains an open question to what extent real-world objects have a fixed capacity limit that is simply higher than that of colors vs. to what extent the concept of "how many things you can remember" is not a valid description of visual working memory (e.g., if the capacity of active storage is largely limited by interference between items, then there is no fixed limit).

There are many important differences between remembering simple stimuli like colored squares and remembering real-world objects. First, real-world objects connect to conceptual knowledge; second, real-world objects are familiar; and third, they are perceptually more complex than the standard simple stimuli like colored circles or oriented lines. In long-term memory, it has generally been proposed that conceptual knowledge (i.e., meaningfulness) associated with real-world objects is the critical attribute that gives rise to enhanced memory (Bower, Karlin, & Dueck, 1975; Konkle et al., 2010; McWeeny, Young, Hay, & Ellis, 1987). Some existing data are consistent with the hypothesis that conceptual information, rather than complexity itself, is relevant for visual working memory as well. For example, with complex but meaningless objects like 3D cubes, participants perform poorly in memory tasks even with just one or two objects, even with long encoding times (Alvarez & Cavanagh, 2004; Olsson & Poom, 2005), unless ensemble coding processes are used to combine information across objects (Brady & Alvarez, 2015a, 2015b). Thus, perceptual complexity without conceptual meaning results in lower performance than simple stimuli with the same number of objects, whereas for meaningful objects performance is better than simple stimuli. However, the capacity differences for real-world objects relative to complex but meaningless objects could also be due to physical differences in the stimuli.

Thus, in recent work, we have expanded this line of evidence about meaningfulness, showing that even for stimuli that are perceptually identical, there is a major benefit to visual working memory if participants understand these stimuli and can recognize them as meaningful objects. In particular, using Mooney images (black and white two-tone images), we have shown in both long-term memory (Brady, Alvarez, & Stormer, 2019) and visual working memory (Asp, Störmer, & Brady, 2018, 2019) that participants better remember Mooney images in which they perceive the face. That is, even for identical images, being able to recognize something as a meaningful "unit" (a face) rather than treating it as a set of meaningless mid-level features results in improved memory performance. In visual working memory, this results in improved "capacity," and also a larger CDA,

again suggesting a change in the capacity of the active storage system rather than the use of non-active forms of memory.

Together, then, there is significant evidence that visual working memory is importantly different depending on the content of the memory. In particular, stimuli that form meaningful units, or even stimuli that have learned associations (e.g., color pairs), seem to allow for greater performance in visual working memory.

3.5 Expertise and visual working memory

In addition to general knowledge and associations impacting visual working memory, it is often the case in the real world that we must engage working memory in complex tasks for which we have some particular level of expertise. How does visual working memory take into account such person-specific prior knowledge in a particular area?

It is widely known that expertise and knowledge improve our ability to maintain information (long-term working memory; Ericsson & Kintsch, 1995; chunking in chess experts; Gobet et al., 2001; etc.). For example, experts show an increased visual working memory capacity for images in their domain: expert car dealers have a greater capacity to remember cars compared to novices (Curby, Glazek, & Gauthier, 2009).

Recent work has examined the impact of expertise in working memory in other more applied domains of expertise. For example, building on work studying expert radiologists which shows they have increased long-term recognition memory for medical images (Evans et al., 2011), we have recently been examining whether this expertise also impacts their ability to remember mammogram images over shorter, working memory delays (Schill, Wolfe, & Brady, 2019). We have hypothesized that similar to the effects observed in long-term memory, there could be expertise-specific effects in working memory not because of the well-known benefits of chunking (e.g., Cowan, 2001; Gobet et al., 2001), but because of improvement that occurs because of existing memory for what variation exists for an image in an expert's domain (e.g., they know how mammograms might vary and so can encode diagnostic features).

Consistent with this idea, one line of recent work has claimed that expertise effects in visual working memory may arise in particular from enhanced consolidation of items into memory—that is, from the ability to encode more information more quickly, rather than a change in the total limit on what can be encoded (Xie & Zhang, 2018). In particular, by looking

at expert Pokemon players, who have knowledge of some particular individual Pokemon but not others (because of the time period of their introduction), Xie and Zhang have found across a number of papers this consolidation effect (Xie & Zhang, 2017a) as well as enhanced capacity for stimuli in experts (Xie & Zhang, 2017b).

Future studies that delve further into the relationship between visual working memory and expertise will enable researchers to look at how expertise develops over time, and to test ways to expedite the process of becoming an expert. Ultimately, this could lead to novel advancements in the literature on both applied and basic theories of memory, as well as new methodology to predict and enhance cognitive performance in real-world tasks.

Overall, we believe there is significant evidence that working memory is impacted by existing knowledge, both in terms of the general meaningfulness of the stimuli as well as person-specific expertise. How might such conceptual knowledge or other crystallized long-term memories enhance active storage in working memory? One hypothesis is that active maintenance is limited by interference between the neural populations that must be held active (e.g., Cohen, Konkle, Rhee, Nakayama, & Alvarez, 2014). Because having more crystallized knowledge about an object or object category results in more relevant neural populations to support memory, attentional mechanisms can maintain more active information successfully when dealing with meaningful objects (Wyble, Swan, & Callahan-Flintoft, 2016). According to this account, neural activity during the storage period of a memory task should be increased for more meaningful stimuli since the active maintenance mechanisms that allow the storage of information successfully maintain more representations of items when the items are meaningful, and thus there are more relevant—and more distinct—neural populations that contain information about these items.

In addition, it may be that the concept of long-term working memory is important (Ericsson & Kintsch, 1995); in particular, that experts and others with significant semantic knowledge partially use working memory to hold in mind retrieval structures that guide their ability to quickly access relevant visual long-term memories.

4. Conclusion

Both visual attention and visual working memory have typically been studied most often with simple stimuli and low-level paradigms. There are significant benefits to this approach; in particular, using simple stimuli allows

us to understand the representations of the stimuli, because we have some understanding of basic visual processing and of the representations of simple stimuli in the visual system at these levels. However, studying the world with simplified stimuli, made up of discrete, single-feature objects on blank backgrounds, leaves much left unstudied about how attention and visual working memory function in more realistic settings. We have argued here that visual attention and visual working memory can be studied by scaling up from the simplest stimuli to more complex scenarios that capture some aspects of real-world settings (e.g., scene structure instead of just segmented objects, meaningful stimuli, auditory inputs), while still remaining in the realm of well-controlled stimuli and well-understood tasks. We believe this approach, which we have been taking in our labs, will allow a more generalizable set of knowledge about visual attention and visual working memory.

References

Aagten-Murphy, D., & Bays, P. M. (2018). Functions of memory across saccadic eye movements. *Current Topics in Behavioral Neurosciences*.

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*(1), 20–29.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*(2), 106–111.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, *19*(4), 392.

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics: Efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 7345.

Andersen, S. K., Hillyard, S. A., & Müller, M. M. (2013). Global facilitation of attended features is obligatory and restricts divided attention. *Journal of Neuroscience*, *33*(46), 18200–18207.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*(2), 157–162.

Arman, A. C., Ciaramitaro, V. M., & Boynton, G. M. (2006). Effects of feature-based attention on the motion aftereffect at remote locations. *Vision Research*, *46*(18), 2968–2976. https://doi.org/10.1016/j.visres.2006.03.003.

Asp, I., Störmer, V., & Brady, T. F. (2018). *Perceptually-matched images that are meaningful are remembered better and result in increased CDA in visual working memory*. San Diego, CA: Poster presented at the Society for Neuroscience.

Asp, I. E., Störmer, V. S., & Brady, T. F. (2019). Greater visual working memory capacity for visually-matched stimuli when they are recognized as meaningful. *PsyArXiv Preprint*. https://doi.org/10.31234/osf.io/r6njf.

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science*, *18*(7), 622. Blackwell Publishing Ltd.

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29. https://doi.org/10.1146/annurev-psych-120710-10042.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*(12), 13.

Baldauf, D., & Desimone, R. (2014). Neural mechanisms of object-based attention. *Science*, *344*(6182), 424–427.

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 7.

Boehler, C. N., Schoenfeld, M. A., Heinze, H.-J., & Hopf, J.-M. (2011). Object-based selection of irrelevant features is not confined to the attended object. *Journal of Cognitive Neuroscience*, *23*(9), 2231–2239. https://doi.org/10.1162/jocn.2010.21558.

Boutonnet, B., & Lupyan, G. (2015). Words jump-start vision: A label advantage in object recognition. *Journal of Neuroscience*, *35*(25), 9329–9335.

Bower, G. H., Karlin, M. B., & Dueck, A. (1975). Comprehension and memory for pictures. *Memory & Cognition*, *3*(2), 216–220. https://doi.org/10.3758/BF03212900.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384. SAGE Publications.

Brady, T. F., & Alvarez, G. A. (2015a). No evidence for a fixed object limit in working memory: Spatial ensemble representations inflate estimates of working memory capacity for complex objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 921–929.

Brady, T. F., & Alvarez, G. A. (2015b). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of Vision*, *15*(15), 6.

Brady, T. F., & Chun, M. M. (2007). Spatial constraints on learning in visual search: Modeling contextual cueing. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 798–815.

Brady, T. F., Alvarez, G., & Störmer, V. (2019). The role of meaning in visual memory: Face-selective brain activity predicts memory for ambiguous face stimuli. *Journal of Neuroscience*, *39*(6), 1100–1108.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*(4), 487–502.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, *11*(5), 4.

Brady, T. F., Konkle, T., Oliva, A., & Alvarez, G. A. (2009). Detecting changes in real-world objects: The relationship between visual long-term memory and change blindness. *Communicative & Integrative Biology*, *2*(1), 1–3.

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.

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

Brady, T. F., & Tenenbaum, J. B. (2010). Encoding higher-order structure in visual working memory: A probabilistic model. In *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 411–416).

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

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525.

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 753.

Chapman, A. F., & Störmer, V. S. (2018). *Feature-based attention is not confined by object boundaries: Spatially global enhancement of irrelevant features*. PsyArxiv Preprint. https://doi.org/10.31234/osf.io/356vk.

Chen, Y. C., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition*, *114*(3), 389–404.

Chen, Y. C., & Spence, C. (2011). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(5), 1554.

Chen, D., Yee Eng, H., & Jiang, Y. V. (2006). Visual working memory for trained and novel polygons. *Visual Cognition*, *14*(1), 37–54. https://doi.org/10.1080/13506280544000282.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*(4), 393–404.

Chooi, W.-T., & Thompson, L. a. (2012). Working memory training does not improve intelligence in healthy young adults. *Intelligence*, *40*(6), 531–542. https://doi.org/10.1016/j.intell.2012.07.004.

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

Chunharas, C., Rademaker, R. L., Brady, T. F., & Serences, J. (2019). *Adaptive distortions in visual working memory*. PsyArXiv. Preprint. *https://doi.org/10.31234/osf.io/e3m5a*.

Chunharas, C., Rademaker, R. L., Sprague, T. C., Brady, T. F., & Serences, J. (2019). Separating memoranda in depth increases visual working memory performance. *Journal of Vision*, *19*, 4.

Ciaramitaro, V. M., Mitchell, J. F., Stoner, G. R., Reynolds, J. H., & Boynton, G. M. (2011). Object-based attention to one of two superimposed surfaces alters responses in human early visual cortex. *Journal of Neurophysiology*, *105*(3), 1258–1265. https://doi.org/10.1152/jn.00680.2010.

Cohen, M. A., Konkle, T., Rhee, J., Nakayama, K., & Alvarez, G. A. (2014). Processing multiple visual objects is limited by overlap in neural channels. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(24), 8955–8960.

Cohen, E. H., & Tong, F. (2015). Neural mechanisms of object-based attention. *Cerebral Cortex*, *25*(4), 1080–1092.

Compte, A., & Wang, X. J. (2005). Tuning curve shift by attention modulation in cortical neurons: A computational study of its mechanisms. *Cerebral Cortex*, *16*(6), 761–778.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, *104*(2), 163.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Cowan, N. (2005). *Working memory capacity*. Psychology Press.

Cukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, *16*(6), 763.

Curby, K. M., Glazek, K., & Gauthier, I. (2009). A visual short-term memory advantage for objects of expertise. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(1), 94.

David, S. V., Hayden, B. Y., Mazer, J. A., & Gallant, J. L. (2008). Attention to stimulus features shifts spectral tuning of V4 neurons during natural vision. *Neuron*, *59*(3), 509–521.

de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *The Quarterly Journal of Experimental Psychology*, *62*(9), 1716–1722.
Dufour, A. (1999). Importance of attentional mechanisms in audiovisual links. *Experimental Brain Research*, *126*(2), 215–222.
Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues. *Cognition*, *143*, 93–100.
Endress, A. D., & Potter, M. C. (2014). Large capacity temporary visual memory. *Journal of Experimental Psychology. General*, *143*(2), 548–565. https://doi.org/10.1037/a0033934.
Eng, H. Y., Chen, D., & Jiang, Y. V. (2005). Visual working memory for simple and complex visual stimuli. *Psychonomic Bulletin & Review*, *12*(6), 1127–1133.
Epstein, R. (2005). The cortical basis of visual scene processing. *Visual Cognition*, *12*(6), 954–978.
Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601.
Ericsson, K., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211–245. https://doi.org/10.1037/0033-295X.102.2.211.
Ernst, Z. R., Boynton, G. M., & Jazayeri, M. (2013). The spread of attention across features of a surface. *Journal of Neurophysiology*, *110*(10), 2426–2439. https://doi.org/10.1152/jn.00828.2012.
Evans, K. K., Cohen, M. A., Tambouret, R., Horowitz, T., Kreindel, E., & Wolfe, J. M. (2011). Does visual expertise improve visual recognition memory? *Attention, Perception, & Psychophysics*, *73*(1), 30–35.
Evans, K. K., & Treisman, A. (2005). Perception of objects in natural scenes: Is it really attention free? *Journal of Experimental Psychology: Human Perception and Performance*, *31*(6), 1476.
Feng, W., Störmer, V. S., Martinez, A., McDonald, J. J., & Hillyard, S. A. (2014). Sounds activate visual cortex and improve visual discrimination. *Journal of Neuroscience*, *34*(29), 9817–9824.
Florey, J., Clifford, C. W., Dakin, S., & Mareschal, I. (2016). Spatial limitations in averaging social cues. *Scientific Reports*, *6*, 32210.
Frassinetti, F., Bolognini, N., & Làdavas, E. (2002). Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental Brain Research*, *147*(3), 332–343.
Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity, not quality: The relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review*, *17*(5), 673–679.
Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience*, *25*(20), 5004–5012.
Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, *11*(5), 473–490.
Gobet, F., Lane, P. C., Croker, S., Cheng, P. C., Jones, G., Oliver, I., et al. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*(6), 236–243.
Greene, M., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, *20*(4), 464–472.
Haberman, J., Brady, T., & Alvarez, G. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General*, *144*(2), 432–446.
Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), R751.
Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 718.

Hollingworth, A. (2004). Constructing visual representations of natural scenes: The roles of short-and long-term visual memory. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(3), 519.

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., & Suzuki, S. (2008). Characteristic sounds facilitate visual search. *Psychonomic Bulletin & Review*, *15*(3), 548–554.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(19), 6829–6833. https://doi.org/10.1073/pnas.0801268105.

Jiang, Y. V., Olson, I. R., & Chun, M. M. (2000). Organization of visual short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 683–702.

Johnson, J., Spencer, J., Luck, S. J., & Schoner, G. (2009). A dynamic neural field model of visual working memory and change detection. *Psychological Science*, *20*(5), 568–577.

Katzner, S., Busse, L., & Treue, S. (2009). Attention to the color of a moving stimulus modulates motion-signal processing in macaque area MT: Evidence for a unified attentional system. *Frontiers in Systems Neuroscience*, *3*, 12. https://doi.org/10.3389/neuro.06.012.2009.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558.

Kravitz, D. J., Saleem, K. S., Baker, C. I., & Mishkin, M. (2011). A new neural framework for visuospatial processing. *Nature Reviews Neuroscience*, *12*(4), 217.

Lau, J. S. H., & Brady, T. F. (2018a). Ensemble statistics accessed through proxies: Range heuristic and dependence on low-level properties in variability discrimination. *Journal of Vision*, *18*(9), 3.

Lau, J. S. H., & Brady, T. F. (2018b). *Intuitive physics under cognitive load: Multiple object tracking benefits from realistic physics*. PsyArXiv Preprint. https://doi.org/10.31234/osf.io/6t859.

Lee, D. K., Itti, L., Koch, C., & Braun, J. (1999). Attention activates winner-take-all competition among visual filters. *Nature Neuroscience*, *2*(4), 375.

Lescroart, M. D., & Gallant, J. L. (2019). Human scene-selective areas represent 3D configurations of surfaces. *Neuron*, *101*(1), 178–192.e7.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(14), 9596–9601.

Lin, P., & Luck, S. (2009). The influence of similarity on visual working memory representations. *Visual Cognition*, *17*(3), 356–372.

Lin, P., & Luck, S. (2012). Proactive interference does not meaningfully distort visual working memory capacity estimates in the canonical change detection task. *Frontiers in Psychology*, *3*, 42.

Ling, S., Liu, T., & Carrasco, M. (2009). How spatial and feature-based attention affect the gain and tuning of population responses. *Vision Research*, *49*(10), 1194–1204. https://doi.org/10.1016/j.visres.2008.05.025.

Liu, T., Larsson, J., & Carrasco, M. (2007). Feature-based attention modulates orientation-selective responses in human visual cortex. *Neuron*, *55*(2), 313–323. https://doi.org/10.1016/j.neuron.2007.06.030.

Liu, T., Slotnick, S. D., Serences, J. T., & Yantis, S. (2003). Cortical mechanisms of feature-based attentional control. *Cerebral Cortex*, *13*(12), 1334–1343. https://doi.org/10.1093/cercor/bhg080.

Liu, T., Stevens, S. T., & Carrasco, M. (2007). Comparing the time course and efficacy of spatial and feature-based attention. *Vision Research*, *47*(1), 108–113. https://doi.org/10.1016/j.visres.2006.09.017.

Lu, Z. L., & Dosher, B. A. (2004). Spatial attention excludes external noise without changing the spatial frequency tuning of the perceptual template. *Journal of Vision*, *4*(10), 10.
Luck, S. J. (2008). Visual short-term memory. In S. J. Luck & A. Hollingworth (Eds.), *Visual Memory* (pp. 43–85). New York: Oxford University Press.
Luck, S. J., Chelazzi, L., Hillyard, S. A., & Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of Neurophysiology*, 77(1), 24–42.
Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281.
Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General*, *141*(1), 170.
Lustig, A. G., & Beck, D. M. (2012). Task-relevant and task-irrelevant dimensions are modulated independently at a task-irrelevant location. *Journal of Cognitive Neuroscience*, *24*(9), 1884–1895. https://doi.org/10.1162/jocn_a_00249.
Martinez-Trujillo, J. C., & Treue, S. (2004). Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current Biology*, *14*(9), 744–751. https://doi.org/10.1016/j.cub.2004.04.028.
McAdams, C. J., & Maunsell, J. H. (2000). Attention to both space and feature modulates neuronal responses in macaque area V4. *Journal of Neurophysiology*, *83*(3), 1751–1755.
McDonald, J. J., Störmer, V. S., Martinez, A., Feng, W., & Hillyard, S. A. (2013). Salient sounds activate human visual cortex automatically. *Journal of Neuroscience*, *33*(21), 9194–9201.
McDonald, J. J., Teder-Sälejärvi, W. A., & Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception. *Nature*, *407*(6806), 906.
Mcdonald, J. J., Teder-Sälejärvi, W. A., Russo, F. D., & Hillyard, S. A. (2003). Neural substrates of perceptual enhancement by cross-modal spatial attention. *Journal of Cognitive Neuroscience*, *15*(1), 10–19.
McWeeny, K., Young, A., Hay, D., & Ellis, A. (1987). Putting names to faces. *British Journal of Psychology*, *78*, 143–146.
Melcher, D. (2001). Persistence of visual memory for scenes. *Nature*, *412*(6845), 401. https://doi.org/10.1038/35086646.
Melcher, D. (2006). Accumulation and persistence of memory for natural scenes. *Journal of Vision*, 6(*1*), 8.
Melcher, D., Papathomas, T. V., & Vidnyánszky, Z. (2005). Implicit attentional selection of bound visual features. *Neuron*, *46*(5), 723–729. https://doi.org/10.1016/j.neuron.2005.04.023.
Meredith, M. A., & Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science*, *221*(4608), 389–391.
Mishra, J., Martinez, A., Sejnowski, T. J., & Hillyard, S. A. (2007). Early cross-modal interactions in auditory and visual cortex underlie a sound-induced visual illusion. *Journal of Neuroscience*, *27*(15), 4120–4131.
Mitchell, J. F., Sundberg, K. A., & Reynolds, J. H. (2007). Differential attention-dependent response modulation across cell classes in macaque visual area V4. *Neuron*, *55*(1), 131–141.
Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, *229*(4715), 782–784.
Morey, C. C., Cong, Y., Zheng, Y., Price, M., & Morey, R. D. (2015). The color-sharing bonus: Roles of perceptual organization and attentive processes in visual working memory. *Archives of Scientific Psychology*, *3*(1), 18.
Motter, B. C. (1994). Neural correlates of attentive selection for color or luminance in extrastriate area V4. *Journal of Neuroscience*, *14*(4), 2178–2189.

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics*, *70*(5), 772–788.

Nastase, S. A., Connolly, A. C., Oosterhof, N. N., Halchenko, Y. O., Guntupalli, J. S., Visconti di Oleggio Castello, M., et al. (2017). Attention selectively reshapes the geometry of distributed semantic representation. *Cerebral Cortex*, *27*(8), 4277–4291.

Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition*, *128*(1), 56–63.

O'Craven, K. M., Downing, P. E., & Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature*, *401*, 584–587. https://doi.org/10.1038/44134.

O'Craven, K. M., Rosen, B. R., Kwong, K. K., Treisman, A., & Savoy, R. L. (1997). Voluntary attention modulates fMRI activity in human MT-MST. *Neuron*, *18*(4), 591–598. https://doi.org/10.1016/S0896-6273(00)80300-1.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3), 145–175. Springer.

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. Elsevier.

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, *13*(11), 4700–4719.

Olson, I. R., & Jiang, Y. (2004). Visual short-term memory is not improved by training. *Memory & Cognition*, *32*(8), 1326–1332.

Olson, I. R., Jiang, Y., & Moore, K. S. (2005). Associative learning improves visual working memory performance. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(5), 889.

Olsson, H., & Poom, L. (2005). Visual memory needs categories. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(24), 8776–8780. https://doi.org/10.1073/pnas.0500810102.

Orhan, A. E., & Jacobs, R. A. (2013). A probabilistic clustering theory of the organization of visual short-term memory. *Psychological Review*, *120*(2), 297.

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

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, *4*, 739.

Peelen, M. V., Fei-Fei, L., & Kastner, S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature*, *460*(7251), 94.

Peelen, M. V., & Kastner, S. (2014). Attention in the real world: Toward understanding its neural basis. *Trends in Cognitive Sciences*, *18*(5), 242–250.

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, *13*(1), 25–42.

Potter, M. C. (1993). Very short-term conceptual memory. *Memory & Cognition*, *21*(2), 156–161.

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

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*(3), 179–197.

Reeder, R. R., & Peelen, M. V. (2013). The contents of the search template for category-level search in natural scenes. *Journal of Vision*, *13*(3), 13.

Rosenholtz, R., Huang, J., Raj, A., Balas, B., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision*, *12*, 14.

Sàenz, M., Buracâs, G. T., & Boynton, G. M. (2002). Global effects of feature-based attention in human visual cortex. *Nature Neuroscience*, *5*(7), 631–632. https://doi.org/10.1038/nn876.

Sàenz, M., Buracâs, G. T., & Boynton, G. M. (2003). Global feature-based attention for motion and color. *Vision Research*, 43(6), 629–637. https://doi.org/10.1016/S0042-6989(02)00595-3.

Sanocki, T. (2003). Representation and perception of scenic layout. *Cognitive Psychology*, *47*, 43–86.

Sanocki, T. (2013). Facilitatory priming of scene layout depends on experience with the scene. *Psychonomic Bulletin & Review*, *20*(2), 274–281.

Sanocki, T., & Epstein, W. (1997). Priming spatial layout of scenes. *Psychological Science*, *8*(5), 374–378.

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

Schill, H., Wolfe, J., & Brady, T. F. (2019). In *Memory capacity meets expertise: increased capacity for abnormal images in expert radiologists. Poster to be presented at the Annual meeting of the vision sciences society*.

Schneider, T. R., Engel, A. K., & Debener, S. (2008). Multisensory identification of natural objects in a two-way crossmodal priming paradigm. *Experimental Psychology*, *55*(2), 121–132.

Schoenfeld, M. A., Tempelmann, C., Martinez, A., Hopf, J. M., Sattler, C., Heinze, H. J., et al. (2003). Dynamics of feature binding during object-selective attention. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(20), 11806–11811.

Schurgin, M. W., & Brady, T. F. (2019). When "capacity" changes with set size: Ensemble representations support the detection of across-category changes in visual working memory. *Journal of Vision*, (in press). https://doi.org/10.31234/osf.io/th5rn.

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2018). *Psychophysical scaling reveals a unified theory of visual memory strength*. bioRxiv Preprint https://doi.org/10.1101/325472.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological science*, *5*(4), 195–200.

Serences, J. T. (2016). Neural mechanisms of information storage in visual short-term memory. *Vision Research*, *128*, 53–67.

Serences, J. T., & Boynton, G. M. (2007). Feature-based attentional modulations in the absence of direct visual stimulation. *Neuron*, *55*(2), 301–312. https://doi.org/10.1016/j.neuron.2007.06.015.

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science*, *20*(2), 207–214.

Shafer-Skelton, A., & Brady, T. F. (2019). Scene priming relies primarily on low-level features rather than scene layout. *Journal of Vision*, *19*(1), 14.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions: What you see is what you hear. *Nature*, *408*(6814), 788.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin*, *138*(4), 628.

Sims, C., Jacobs, R., & Knill, D. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, *119*(4), 807–830.

Solomon, J., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision*, *11*(12), 13.

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, *9*(4), 255.

Störmer, V. S., & Alvarez, G. A. (2014). Feature-based attention elicits surround suppression in feature space. *Current Biology*, *24*(17), 1985–1988. https://doi.org/10.1016/j.cub.2014.07.030.

Störmer, V. S., Cohen, M. A., & Alvarez, G., Tuning attention to object categories: Spatially global effects of attention to faces in visual processing, *Journal of Cognitive Neuroscience*, (in press).

Störmer, V. S., Feng, W., Martinez, A., McDonald, J. J., & Hillyard, S. A. (2016). Salient, irrelevant sounds reflexively induce alpha rhythm desynchronization in parallel with slow potential shifts in visual cortex. *Journal of Cognitive Neuroscience*, *28*(3), 433–445.

Störmer, V. S., Li, S.-C., Heekeren, H. R., & Lindenberger, U. (2013). Normal aging delays and compromises early multifocal visual attention during object tracking. *Journal of Cognitive Neuroscience*, *25*(2), 188–202.

Störmer, V. S., McDonald, J. J., & Hillyard, S. A. (2009). Cross-modal cueing of attention alters appearance and early cortical processing of visual stimuli. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(52), 22456–22461.

Sugihara, T., Diltz, M. D., Averbeck, B. B., & Romanski, L. M. (2006). Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex. *Journal of Neuroscience*, *26*(43), 11138–11147.

Swan, G., & Wyble, B. (2014). The binding pool: A model of shared neural resources for distinct items in visual working memory. *Attention, Perception, & Psychophysics*, *76*(7), 2136–2157.

Sweeny, T. D., Haroz, S., & Whitney, D. (2013). Perceiving group behavior: Sensitive ensemble coding mechanisms for biological motion of human crowds. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 329.

Sweeny, T. D., & Whitney, D. (2014). Perceiving crowd attention: Ensemble perception of a crowd's gaze. *Psychological Science*, *25*(10), 1903–1913.

Thompson, T. W., Waskom, M. L., Garel, K. L. A., Cardenas-Iniguez, C., Reynolds, G. O., Winter, R., et al. (2013). Failure of working memory training to enhance cognition or intelligence. *PLoS One*, *8*(5), e63614. https://doi.org/10.1371/journal.pone.0063614.

Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition*, *14*(4–8), 411–443.

Treue, S., & Martinez-Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, *399*, 575–579. https://doi.org/10.1038/21176.

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology*, *71*, 1–26.

Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(5), 1053.

Vetter, P., Smith, F. W., & Muckli, L. (2014). Decoding sound and imagery content in early visual cortex. *Current Biology*, *24*(11), 1256–1262.

Victor, J. D., & Conte, M. M. (2004). Visual working memory for image statistics. *Vision Research*, *44*(6), 541.

Võ, M. L. H., & Henderson, J. M. (2010). The time course of initial scene processing for eye movement guidance in natural scene search. *Journal of Vision*, *10*(3), 14.

Walker, D., & Vul, E. (2014). Hierarchical encoding makes individuals in a group seem more attractive. *Psychological Science*, *25*(1), 230–235.

Wallace, M. T., Ramachandran, R., & Stein, B. E. (2004). A revised view of sensory cortical parcellation. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(7), 2167–2172.

Wannig, A., Rodríguez, V., & Freiwald, W. A. (2007). Attention to surfaces modulates motion processing in extrastriate area MT. *Neuron*, *54*(4), 639–651. https://doi.org/10.1016/j.neuron.2007.05.001.

White, A. L., & Carrasco, M. (2011). Feature-based attention involuntarily and simultaneously improves visual performance across locations. *Journal of Vision*, *11*(6), 1–10. https://doi.org/10.1167/11.6.15.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology*, *69*, 105–129.

Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *2*(5–6), 440–445. https://doi.org/10.1016/S0022-5371(63)80045-6.

Wilken, P., & Ma, W. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 1120–1135. Association for Research in Vision and Ophthalmology.

Wiseman, S., & Neisser, U. (1974). Perceptual organization as a determinant of visual recognition memory. *The American Journal of Psychology*, *87*(4), 675–681.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, *5*(6), 495.

Wyble, B., Swan, G., & Callahan-Flintoft, C. (2016). Measuring visual memory in its native format. *Trends in Cognitive Sciences*, *20*(11), 790–791.

Xie, W., & Zhang, W. (2017a). Familiarity increases the number of remembered Pokémon in visual short-term memory. *Memory & Cognition*, *45*(4), 677–689.

Xie, W., & Zhang, W. (2017b). Familiarity speeds up visual short-term memory consolidation. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(6), 1207–1221.

Xie, W., & Zhang, W. (2018). Familiarity speeds up visual short-term memory consolidation: Electrophysiological evidence from contralateral delay activities. *Journal of Cognitive Neuroscience*, *30*(1), 1–13.

Yamanashi Leib, A. Y., Fischer, J., Liu, Y., Qiu, S., Robertson, L., & Whitney, D. (2014). Ensemble crowd perception: A viewpoint-invariant mechanism to represent average crowd identity. *Journal of Vision*, *14*(8), 26.

Yamanashi Leib, A. Y., Puri, A. M., Fischer, J., Bentin, S., Whitney, D., & Robertson, L. (2012). Crowd perception in prosopagnosia. *Neuropsychologia*, *50*(7), 1698–1707.

Zampini, M., Shore, D. I., & Spence, C. (2005). Audiovisual prior entry. *Neuroscience Letters*, *381*, 217–222.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235.

Zhang, W., & Luck, S. J. (2009). Feature-based attention modulates feedforward visual processing. *Nature Neuroscience*, *12*(1), 24.