What modern vision science reveals about the awareness puzzle: Summary-statistic encoding plus limits on decision complexity underlie the richness of visual perception and its quirky failures

Ruth Rosenholtz, MIT Dept. of Brain & Cognitive Sciences, CSAIL

Abstract

Human beings subjectively experience a rich visual percept. However, when behavioral experiments probe the details of that percept, observers perform poorly, suggesting that vision is impoverished. What can explain this awareness puzzle? Is the rich percept a mere illusion? How does vision work as well as it does? This paper argues for two important pieces of the solution. First, peripheral vision encodes its inputs using a scheme that preserves a great deal of useful information, while losing the information necessary to perform certain tasks. The tasks that are rendered difficult by the peripheral encoding include many of those used to probe the details of visual experience. Second, many tasks used to probe attentional and working memory limits arguably are inherently difficult; poor performance on these tasks may indicate limits on decision-making capacity. Together, these two components can explain a wide variety of phenomena, including vision's marvelous successes, its quirky failures, and our rich subjective impression of the visual world.

1. Introduction

Without loss of generality, we can assume that all perception arises from performing some visual task, i.e. from making some inference about the visual world based on the available information, working memory, and prior knowledge. Limits clearly exist, at any given moment, in terms of both the information available and the tasks one can successfully

perform. Visual awareness is likely even more limited; organisms can carry out considerable visual processing without awareness (Helmholtz, 1867; Koch & Crick, 2001). Because of these limits on visual processing and awareness, real-world vision involves an iterative process. We start with some – possibly unconscious – task, i.e. some question about the world. For instance, we might start by asking, "what is the layout of this room?" We do our best to complete that task. If necessary we can gain more information by taking actions such as moving our eyes. In the next instance, we shift to another task to gain more understanding of the visual world. For example, we might next query, "are there any people here?" Similarly, our awareness of what we know about the visual world shifts from moment to moment.

In studying this process of understanding and becoming aware of the visual world, a fundamental puzzle has arisen. On one hand, we subjectively experience a rich visual world, effortlessly perceived (Dennett, 1991; Noë, 2002). However, when probed on the details, observers know surprisingly little. The rich experience suggests a highly capable visual system, whereas poor performance reporting details suggests that perception is impoverished. For the purposes of this paper, I refer to this puzzling combination of rich subjective experience and poor objective task performance as "the awareness puzzle," though of course it is far from the only puzzle when it comes to understanding awareness (Tononi, Boly, Massimini, & Koch, 2016).

For example, we subjectively experience a rich awareness of a real-world scene (Dennett, 1991). However, change a portion of that scene (while masking transients that would provide a cue), and observers have difficulty noticing what changed (e.g. (Rensink, O'Regan, & Clark, 1997)). Similarly, while we experience a rich percept of an ensemble of similar items, observers perform poorly when asked to report the features of a particular item (Ariely, 2001; Chong & Treisman, 2005; Haberman & Whitney, 2009). Furthermore, it

is difficult to search for a particular target item unless that item has a distinct basic feature like orientation, color, or motion. Search can be difficult even when, upon examination, target and distractors appear quite distinct, e.g. when searching for a T among Ls. Difficult search, then, suggests that the details that distinguish the search items must be unavailable; otherwise search would be easy.

An influential theory (Feature Integration Theory, or FIT, (Treisman & Gelade, 1980)) proposed that poor search performance arises from a particular kind of limited capacity: limited access to higher-level processing. According to this theory, observers can quickly and easily perform tasks that require only basic feature maps; such tasks rely only on "preattentive" visual processing. However, any tasks which rely on binding or conjoining an object's features, such as distinguishing a T from an L, require selective attention. According to this theory, attention serially selects what information travels through the limited capacity channel to receive higher-level processing.

Attention, in turn, appears to have greatly limited capacity (Figure 1). Multiple object tracking (MOT) tasks, for instance, suggest that observers can attend to and track only about 4 objects at a time (e.g. (Pylyshyn & Storm, 1988)). Furthermore, there is often a cost to performing more than one task at once (VanRullen, Reddy, & Koch, 2004), particularly when one of the tasks is unknown to the observer, as in the phenomenon of inattentional blindness (Mack & Rock, 1998).

If perception is poor without attention, and attention has limited capacity, then at a given instant we cannot perceive very much. Furthermore, we cannot merely build up a rich percept by rapidly shifting attention and remembering what we have previously perceived, because visual working memory itself appears to have a low capacity of approximately 4 items (e.g. (Luck & Vogel, 1997)).

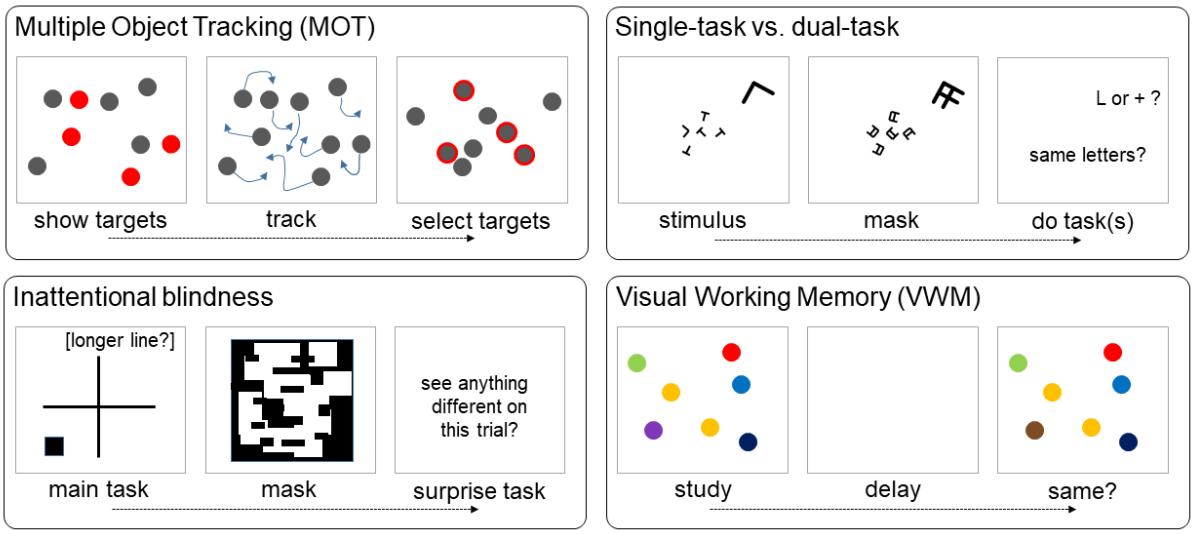

Figure 1. Tasks such as visual search, change detection, and perception of individual items of a set have suggested that perception is limited without attention. Furthermore, paradigms shown here, such as multiple object tracking, dual-task, and inattentional blindness, have suggested that attention is limited. Visual working memory tasks, in turn, have suggested that memory has limited capacity. In each paradigm depicted, time advances to the right, as indicated by the arrow. This paper argues that these tasks are inherently difficult.

A number of philosophers and vision researchers have noted the confusing collection of phenomenology described above, and have proposed theories to address the underlying puzzles. The first two theories are philosophical in nature, attempting to make sense of the apparent contradiction between the rich subjective experience and poor performance at a number of objective tasks. The second two classes of theory, more vision science than philosophy, suggest visual mechanisms to account for both the awareness puzzle and for real-world vision.

The first philosophical theory, here referred to as the *illusion theory*, suggests that the rich subjective impression is merely an illusion, and therefore not incompatible with the impoverished perception observed in behavioral experiments (O'Regan, 1992; Rensink, O'Regan, & Clark, 1997; Blackmore, Brelstaff, Nelson, & Troscianko, 1995; Dennett, 1991; Dennett, 1998). This illusion theory must contend with empirical evidence in favor of an objectively richer percept. Observers can rapidly get the gist of a scene (e.g., (Potter, 1975; Loftus & Ginn, 1984; Rousselet, Joubert, & Fabre-Thorpe, 2005; Loschky, et al., 2007;

Greene & Oliva, 2009; Potter & Fox, 2009), and this gist includes rich information about that scene (Fei-Fei, Iyer, Koch, & Perona, 2007). Similarly, we can rapidly extract properties of an ensemble (Chong & Treisman, 2003; Chong & Treisman, 2005; Ariely, 2001; Alvarez, 2011; Haberman & Whitney, 2009). Clearly these results are, at minimum, problematic for the original Feature Integration Theory, as noted in (Treisman, 2006), though it remains unclear whether the details objectively available to observers suffice to explain the subjective experience.

The second philosophical theory posits that we are aware of more than we can act upon (Lamme, 2010; Block, 2011). In this theory, here referred to as the *inaccessibility theory*, the rich percept is real, but the information is perversely inaccessible when it comes to making decisions or otherwise taking action. Proponents of this theory need to explain why it appears to lie in opposition to our intuition that awareness should encounter greater limits than perception.

It is not obvious how either philosophical theory would lead to a working visual system. If perceptual richness is mere illusion, how are we so successful at so many visual tasks? Why put energy into awareness but not ensure the ability to act on the available information?

Vision science theories have attempted to account for the awareness puzzle while also explaining how real-world vision might work. One class of theories, for instance, asks how, if preattentive vision is so poor, and attention so limited, can we intelligently shift attention to gather more information? How can we reasonably form and test new hypotheses to gain understanding about the visual world? Suppose I want my coffee mug. To identify it I need to attend to it; where do I direct my attention? It might help me to know that the mug sits on the desk. But this presents a bit of a chicken-and-egg problem: I would have to attend to the desk to identify the desk. If it is my desk, in my office, I might

have prior knowledge of its location. If it is someone else's desk, but I know it is brown, I could use crude preattentive features to filter for brown stuff (Wolfe, Cave, & Franzel, 1989). But what if I know neither piece of information? For that matter, how do I ever perceive task-irrelevant parts of the scene, such as my colleague sitting behind the desk?

Mack and Rock (1998), noting that their inattentional blindness studies seemed to suggest little or no perception without attention, proposed that some information must be capable of "capturing" attention. This suggestion led to the development of a rich literature attempting to uncover the rules of attentional capture. Stimulus-driven, or bottom-up capture could occur if the information is sufficiently "salient" (Theeuwes, 1992), though this might depend upon the task set (Folk, Remington, & Johnston, 1992). Bottom-up saliency, i.e. unusual features, could be computed from the hypothesized preattentive features, e.g. (Itti & Koch, 2001; Rosenholtz, 1999). Capture by salient items could help us notice interesting parts of the scene even if they are not task-relevant. Top-down filters could also reveal task-irrelevant information. For instance, Simons and Chabris (1999) suggested that observers are more likely to notice an unexpected gorilla walking through a basketball game when they count passes of the team with black jerseys, perhaps because the filter for "black" accidentally captures the gorilla, leading to identification. However, taking a step back, attentional capture seems like an odd proposal for how vision might work: the visual system makes up for poor preattentive processing both by being easily distracted by irrelevant salient stuff, and by having top-down filters that accidentally capture task-irrelevant items with crude low-level similarity to the task-relevant items. This is no way to design a visual system, and it seems unlikely that such capture can explain vision's successes (Rosenholtz, Huang, & Ehinger, 2012; Nakayama, 1990).

A second class of vision theories has built on the observation that classic selective attention theory can account for some of vision's quirky failures, but is problematic when it

comes to explaining vision's marvelous successes. This might suggest that the visual system augments the selective attention pathway with additional information. Scenes and sets, for example, might be processed in a separate, non-selective pathway (Wolfe, Vo, Evans, & Greene, 2011; Cohen, Dennett, & Kanwisher, 2016). Alternatively, different modes of attention might make available different information; diffusely attend to a scene as a whole and get the gist, or attend to a set of items and gain access to ensemble properties like the mean size (Nakayama, 1990; Rensink, 2001; Treisman, 2006). Both mechanisms (separate pathway and diffuse attentional mode) are presumed to utilize a different sort of encoding, unlike that for ordinary object recognition (Nakayama, 1990; Treisman, 2006; Wolfe, Vo, Evans, & Greene, 2011; Cohen, Dennett, & Kanwisher, 2016). Researchers have suggested that mechanisms might encode some sort of summary statistics that would support both scene and ensemble tasks (Treisman, 2006; Oliva & Torralba, 2006; Haberman & Whitney, 2011; Wolfe, Vo, Evans, & Greene, 2011; Cohen, Dennett, & Kanwisher, 2016).

I will argue that the tasks that show limits are inherently difficult tasks, given the representation of information in the visual system. I propose a unified theory which explains the rich subjective experience, the apparent limits, and the power of real-world vision. In other words, looked at in the right way, there is in fact no awareness puzzle. Various researchers have suggested or echoed aspects of this theory (Ariely, 2001; Treisman, 2006; Oliva & Torralba, 2006; Balas, Nakano, & Rosenholtz, 2009; Freeman & Simoncelli, 2011; Rosenholtz, 2011; Haberman & Whitney, 2011; Wolfe, Vo, Evans, & Greene, 2011; Rosenholtz, Huang, & Ehinger, 2012; Cohen, Dennett, & Kanwisher, 2016). Part of the goal of this paper is to clarify and/or modernize these previous suggestions, examine the issues that remain, and suggest a new piece of the puzzle. In particular, the next section suggests that change blindness and visual search phenomena may arise in large part from an efficient encoding in peripheral vision. This has significant implications

for the degree to which vision is impoverished. Section 3.1 further proposes that the tasks purporting to demonstrate the low capacity of attention and working memory may be inherently difficult tasks. This implies both that we as a field may have misjudged how impoverished vision is, and that visual and cognitive processing may encounter capacity limits on decision-making processes. Section 3.2 fleshes out a proposed limit on decision complexity, and its implications for the rich subjective experience and real-world vision.

2. An efficient encoding in peripheral vision explains many of the puzzles of vision

2.1 Change blindness and difficult search may be due to the limits of peripheral vision, not limits on attention

Change blindness refers to the difficulty detecting a change to an image or scene. In the lab, the experimental paradigm often involves flickering between two versions of an image, while introducing a brief blank frame between the pair in order to disrupt motion cues (Rensink, O'Regan, & Clark, 1997). The phenomenon is related to the childhood puzzle in which one must find the differences between two side-by-side images (Scott-Brown, Baker, & Orbach, 2000).

Many researchers have interpreted change blindness as probing the limits of perception or memory without attention, e.g. (Rensink, O'Regan, & Clark, 1997; Hollingworth & Henderson, 2002; O'Regan, Rensink, & Clark, 1999; O'Regan, 1992; Scholl, 2000). Supposedly, the observer manipulates a spotlight of attention, and perception is richer within that spotlight than outside of it. The difficulty of detecting a change appears to imply that little perception occurs without attention.

However, others have suggested that change blindness might be due in part to peripheral vision. (Here I use the term to mean visual processing that occurs in the part of the visual field outside the foveola.) Peripheral vision is known to be poor relative to foveal

vision; visual acuity, contrast sensitivity, color vision, and motion perception all vary with eccentricity, i.e. with distance from the center of gaze (see (Rosenholtz, 2016) for a recent review). A more consequential difference concerns peripheral vision's degradation in the presence of clutter, known as crowding. The phenomenon of visual crowding illustrates that loss of information in the periphery is not merely due to reduced acuity. In classic demonstrations, a target letter is easily identified when presented in the periphery on its own, but becomes difficult to recognize when flanked too closely by other stimuli, such as other letters. An observer might see the crowded letters in the wrong order, they might not see the target at all, or they might see a confusing jumble of shapes made up of parts from multiple letters (Lettvin, 1976). Crowding occurs with a broad range of stimuli (see (Pelli & Tillman, 2008) for a review). It need not involve an individuated "target" and "flankers, *per se*, but rather can occur in peripheral perception of complex objects and scenes (Martelli, Majaj, & Pelli, 2005). Moreover, the degree of difficulty an observer has in making sense of peripheral stimuli varies considerably with the stimulus and task (Andriessen & Bouma, 1976; Kooi, Toet, Tripathy, & Levi, 1994; Livne & Sagi, 2007; Sayim, Westheimer, & Herzog, 2010; Manassi, Sayim, & Herzog, 2012).

At any given moment during a change-detection experiment, the changed region likely lies in the peripheral visual field. This raises the question of whether observers have difficulty detecting changes not because of limited perception without attention, but rather because of limited perception in peripheral vision. Henderson and Hollingworth (1999) showed that observers are more likely to detect the change once they have fixated on or near that change. O'Regan et al. (2000) found that probability of detection depends upon the distance between the initial fixation and the change. Furthermore, research from Parker (1978) and Zelinsky (2001) has suggested that observers can notice the change in the periphery, and that salient changes can even be detected without fixation.

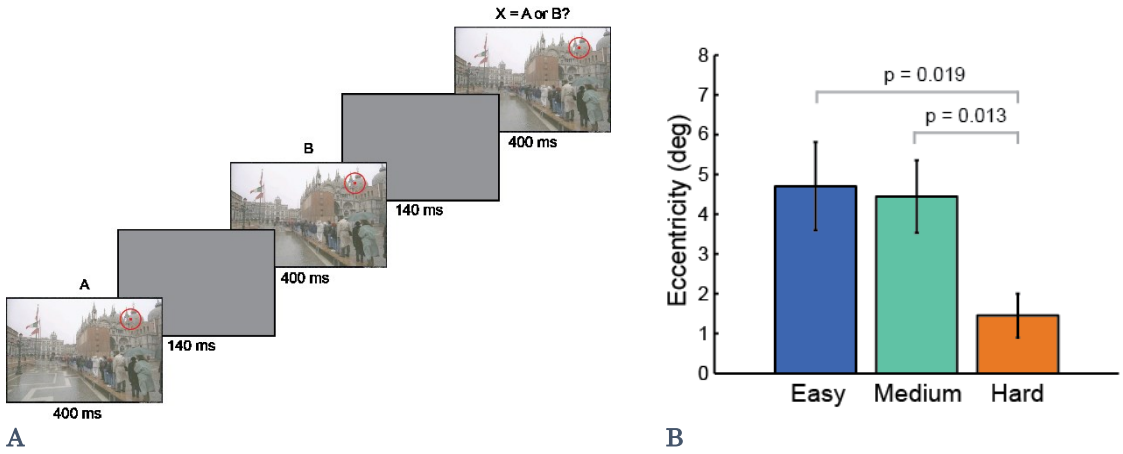

Figure 2. Peripheral vision is a factor in change blindness. A. Observers discriminated known changes in an A-B-X paradigm that requires them to identify whether the final image matches the first or the second image in the sequence. Fixation was enforced at various distances to the change. The red circle shows one such fixation. The difference for this sequence was in the pattern on the ground. B. The threshold eccentricity (distance to the change) for easy-, medium-, and hard-to-detect changes. Harder changes require closer fixation in order to be discriminated.

We have found additional evidence that peripheral vision is a factor in change blindness. We first measured observer change detection performance for a number of image pairs, using a standard flicker paradigm (Rensink, O'Regan, & Clark, 1997). Based on this data, we categorized these standard change blindness stimuli as easy, medium and hard, according to the time needed to detect the changes. We then measured difficulty detecting a known and presumably attended change using peripheral vision (Figure 2A). We found that for the hard changes, observers needed to fixate significantly closer to the change in order to perceive it (Figure 2B), even though they knew in advance the identity of the change and its location (Smith, Sharan, Park, Loschky, & Rosenholtz, under revision). Changes that are harder to detect in a flicker paradigm are harder to see in the periphery, even when observers know the change and its location, and presumably attend to the change. These results suggest a more tenuous connection between change blindness and attentional limits. Furthermore, they suggest that peripheral vision helps detect changes; otherwise peripheral discriminability would not predict change detection difficulty. Change detection may occur across the visual field, in parallel, though the observer may not be aware of

looking for changes in the periphery. As we will see, this paradigm shift in thinking about change blindness has significant implications for the awareness puzzle.

We can similarly reexamine visual search. In the traditional view, search experiments probe limits of attention. By comparing conditions that lead to difficult vs. easy visual search, we supposedly determine at what stage selection occurs, and what processing is preattentive. Experiments have generally shown that search is difficult whenever distinguishing the search target from other distractor items requires more than a simple basic feature like color or motion. This implies that only basic features – often referred to as "feature maps" – can be computed preattentively, and selection occurs early in visual processing. Because attention is a limited resource, this implies that vision is highly impoverished.

However, considerable research has suggested that peripheral vision plays a significant role in search difficulty. If so, peripheral vision at minimum acts as a confound in most search experiments. Carrasco and colleagues have found eccentricity effects in search, and shown that search performance changes when one m-scales the stimuli to reduce peripheral factors (Carrasco, Evert, Chang, & Katz, 1995; Carrasco & Frieder, 1997; Carrasco & Yeshurun, 1998; Carrasco, McLean, Katz, & Frieder, 1998). Peripheral discriminability of Gabors in noise predicts search for Gabor targets (Geisler, Perry, & Najemnik, 2006). There have also been hints that search difficulty stems from crowding in peripheral vision (Erkelens & Hooge, 1996; Gheri, Morgan, & Solomon, 2007).

As in the case of change blindness, we have extended this work on search and crowding by having observers attend to the periphery, and perform peripheral discrimination of a crowded target-present from a target-absent patch. We have shown that this peripheral discriminability predicts search performance (Figure 3). Importantly, many of the classic phenomena that originally motivated selective attention theory are already

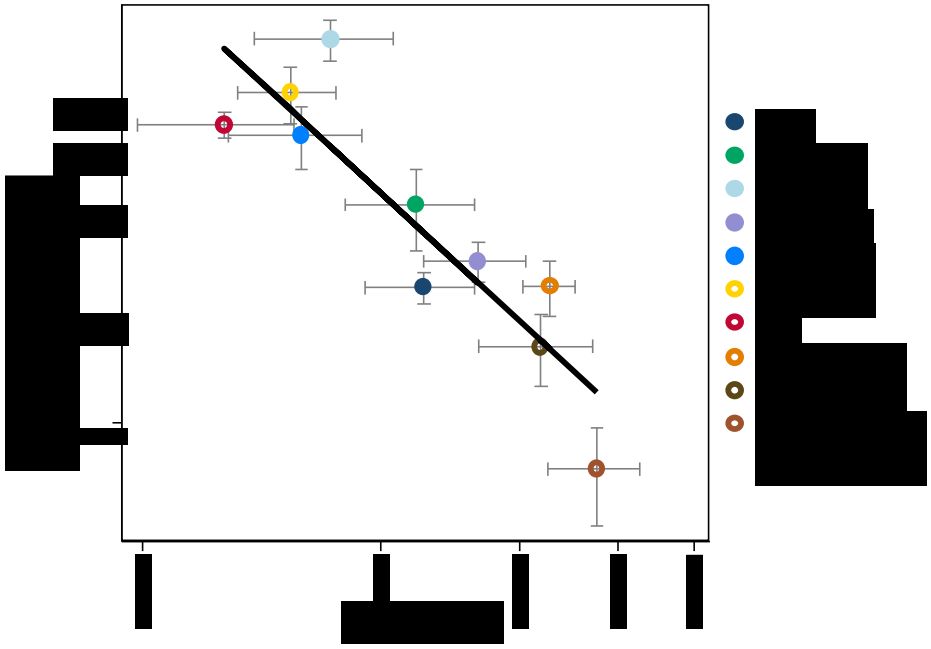

Figure 3. Peripheral discriminability of a crowded target-present vs. target-absent patch (x-axis) predicts search difficulty (y-axis, measured as the slope of the function relating search reaction time to the number of display items). Target-present patches consist of a target flanked by a number of distractors, whereas target-absent patches consist of a distractor flanked by additional distractors. Each symbol represents a different search condition, including both five conditions central to Feature Integration Theory and five problematic conditions showing unexpectedly easy search for a shaded cube among differently shaded cubes. Figure reproduced with permission from (Zhang, Huang, Yigit-Elliot, & Rosenholtz, 2015).

present in peripheral vision under conditions of crowding. Even when an observer attends to the periphery, they have trouble distinguishing a crowded T from a crowded L. They perceive illusory conjunctions. On the other hand, easy search tasks correspond to easy peripheral identification. Peripheral vision preserves the necessary information to identify unique basic features. The strong relationship between search performance and peripheral discriminability, across a wide range of conditions, suggests that search difficulty primarily pinpoints loss of information in peripheral vision, rather than attentional limits or the limits of preattentive processing (Rosenholtz, Huang, Raj, Balas, & Ilie, 2012; Zhang, Huang, Yigit-Elliot, & Rosenholtz, 2015; Chang & Rosenholtz, 2016).

As a result, we argue that neither search nor change blindness clearly support the claim of impoverished vision without attention. Rather, difficulty in those tasks may arise

from limits in peripheral vision. One might ask why this distinction matters, since either explanation implies a loss of information, whether from not attending to a region or from not fixating it. At first glance, either theory would appear to suggest impoverished vision. However, a peripheral vision explanation implies that perception is richer than previously thought. In the attention explanation, unselected stimuli receive virtually no further processing beyond the bottleneck of attention. This means that many, if not most, tasks are impossible without attention. On the other hand, according to the peripheral vision account, difficult change detection and search tasks have relied on information that happens to be lost in peripheral vision; these tasks may be especially difficult, and not imply impoverished vision overall. Peripheral vision preserves a great deal of information, and critically, processing continues. Just what information is preserved, and what tasks that information supports, can best be answered with a model of peripheral vision (Section 2.2).

2.2 A summary statistic encoding in peripheral vision determines difficulty for a range of visual tasks

My lab has argued since 2007 that peripheral vision encodes its inputs in terms of a rich set of image statistics. The term "image statistics" refers to statistics computed over either the pixels of the image or over the outputs of image processing operations such as filters and non-linear operators applied to the image. These statistics are "summary statistics", meaning they pool information over sizeable local regions that grow with the distance to the point of fixation, i.e. the eccentricity. For our candidate model (Balas, Nakano, & Rosenholtz, 2009), we chose as our set of image statistics those from a state-of-the-art model of texture appearance from Portilla & Simoncelli (2000): the marginal distribution of luminance; luminance auto correlation; correlations of the magnitude of responses of oriented V1-like wavelets across differences in orientation, neighboring positions, and scale;

and phase correlation across scale. This seemingly complicated set of parameters is actually fairly intuitive: computing a given second-order correlation merely requires taking responses of a pair of V1-like filters, point-wise multiplying them, and taking the average over the pooling region.

This encoding leads to significant loss of information, and we have accumulated extensive evidence that this loss of information can predict difficulty recognizing peripheral objects in cluttered displays or scenes (Balas, Nakano, & Rosenholtz, 2009; Rosenholtz, Huang, Raj, Balas, & Ilie, 2012; Zhang, Huang, Yigit-Elliot, & Rosenholtz, 2015; Chang & Rosenholtz, 2016; Keshvari & Rosenholtz, 2016; Freeman & Simoncelli, 2011). The loss of information predicts difficult search conditions, while preserving the information necessary to predict easy "popout" search (Rosenholtz, Huang, Raj, Balas, & Ilie, 2012; Zhang, Huang, Yigit-Elliot, & Rosenholtz, 2015; Chang & Rosenholtz, 2016).

In spite of the loss of information that leads to crowding, this encoding preserves a great deal of information. To get a sense of what information is encoded by a rich set of image statistics such as those proposed, one can synthesize images that contain the same statistics but are otherwise random (Rosenholtz, 2011; Freeman & Simoncelli, 2011; Rosenholtz, Huang, & Ehinger, 2012; Ehinger & Rosenholtz, 2016). One should not think of these images as "what the world looks like to peripheral vision." Rather, viewing the synthesized images (e.g. Figure 4), provides intuitions about the information lost and maintained by the peripheral encoding. We have called these syntheses "mongrels", for short. The encoding appears to preserve considerable information about the fact that the underlying image in Figure 4A is a street scene, with people waiting at a bus stop. Detailed information survives about the appearance of the buildings and trees, and about the general layout of the scene. By asking observers to perform scene tasks with these mongrel images, we have demonstrated that the encoding preserves sufficient information to

A B

Figure 4. Information encoded by a rich set of image statistics. A. Original image, theoretical pooling regions superimposed. They grow linearly with eccentricity. B. Image synthesized to have approximately the same local image statistic as the original. This encoding captures a great deal of information, though some of the details are unclear.

quantitatively predict human performance getting the gist of the scene at a glance, including scene category, upcoming turns, presence of a particular object like an animal or a stop sign, and what city appears in the photograph (Rosenholtz, Huang, & Ehinger, 2012; Ehinger & Rosenholtz, 2016).

It is not surprising that the encoding preserves so much useful information, as this scheme involves measuring a large number of image statistics; as many as 1000 per pooling region. This is no mere handful of summary statistics; such an encoding would obviously not support the richness of vision. While little has been done to characterize the information available in our rich subjective impression of the world, it seems plausible that this encoding scheme has sufficient information to support that subjective impression.

Examining Figure 4, however, it is clear that the encoding does not preserve certain details. One cannot read the Thomson Rd. sign, nor easily discriminate the number and types of vehicles. This ambiguity of the details could underlie poor performance in change detection experiments (Freeman & Simoncelli, 2011; Cohen, Dennett, & Kanwisher, 2016; Smith, Sharan, Park, Loschky, & Rosenholtz, under revision). Figure 5 shows a demo of

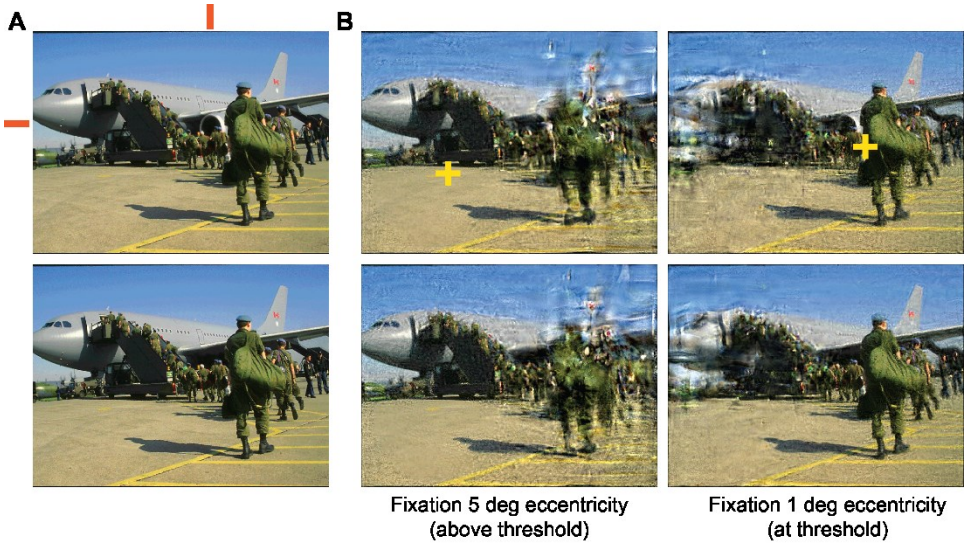

Figure 5. Summary image statistics lose information about the details, which can lead to difficult change detection. A. Image pair. Red bars indicate changed region; the airplane engine present in the upper image but absent in the lower. B. Synthesis visualizes the information available in a summary statistic encoding for a fixation 5 degrees (left) and 1 degree (right) from the change. Note that the change is clear in the latter pair, but not the former.

this same synthesis technique applied to a change detection pair. When fixating 5 degrees away from the change, the model predicts difficulty detecting that change. However when fixating 1 degree away, the change becomes clear, in agreement with our data on discrimination of this change in the periphery (Smith, Sharan, Park, Loschky, & Rosenholtz, under revision).

A summary statistic encoding in peripheral vision, then, seems promising in terms of providing a coherent explanation of a number of diverse phenomena that have previously defied easy explanation. The same encoding predicts relative difficulty of different visual search conditions, as well as scene perception performance. Peripheral vision is clearly a factor in change blindness. While further work (in progress) is necessary to test whether the model can *quantitatively* predict change detection difficulty, results so far appear promising: demonstrations of the information available appear to be in line with difficult change detection (Figure 5), and extensive work, cited above, validates the ability of this encoding to predict peripheral discriminability for a considerable range of conditions. We

have also found evidence that peripheral vision is a factor in some inattentional blindness phenomena (i.e. we have examined the invisible gorilla effects of (Simons & Chabris, 1999); (Rosenholtz, Sharan, & Park, under revision)). Furthermore, a summary statistic encoding in peripheral vision can even make some sense of which tasks are difficult in a dual-task paradigm (Rosenholtz, Huang, & Ehinger, 2012).

2.3 Comparing the proposed encoding scheme to other theories

At this point, it is worth revisiting a few of the previous theories discussed in Section 1. Several of the theories suggested that a different pathway or a different mode of attention might provide information beyond that available in the selective attention pathway/mode, and that the extra information might take the form of some sort of summary statistics (Treisman, 2006; Oliva & Torralba, 2006; Haberman & Whitney, 2011; Wolfe, Vo, Evans, & Greene, 2011; Cohen, Dennett, & Kanwisher, 2016). This proposal should sound like (and in the case of Cohen et al., was at least partially inspired by) our model of peripheral vision. Note, however, that the "extra" pathway alone can explain rather a lot. Researchers added the second pathway to account for good performance on scene and set perception tasks, and in fact a summary statistic encoding does seem promising at predicting performance at those tasks. However, that same encoding can also predict easy vs. difficult search, and likely change blindness; phenomena that allegedly arose from limitations of the selective pathway. What, then, is the purpose of the supposed selective attention pathway, and what are its mechanisms? Our new understanding of peripheral vision demands rethinking attention. Nonetheless, peripheral vision cannot be the whole story; Section 3 suggests an alternative hypothesis to account for additional phenomena not attributable to peripheral vision alone.

The proposed encoding measures a large number of summary *image* statistics, across the field of view, regardless of the contents of the visual stimulus (see also Freeman & Simoncelli (2011), the texture descriptors of Wolfe et al. (2011), and the large number of image statistics hypothesized to underlie the gist of a scene in Oliva & Torralba (2006)). At minimum, a number of previous proposals have lacked clarity on these points. First, summary image statistics are not the same as *ensemble* properties of a set of *items* (Ariely, 2001; Treisman, 2006; Haberman & Whitney, 2011; Cohen, Dennett, & Kanwisher, 2016). Second, a number of researchers have proposed that ensemble properties represent only certain portions of the visual world (Cohen, Dennett, & Kanwisher, 2016), e.g. only sets of similar items (Treisman, 2006; Whitney & Leib, 2018), or only textures (Haberman & Whitney, 2011). Third, a number of previous proposals have implied that the encoding involves only a small number of summary statistics, (e.g. (Cohen, Dennett, & Kanwisher, 2016; Treisman, 2006; Ariely, 2001; Haberman & Whitney, 2011).

Though summary image statistics and ensemble properties of a set of objects are often confused, there exists an important asymmetry between the two: A large set of image statistics cannot only support a variety of scene perception tasks (Ehinger & Rosenholtz, 2016), but also plausibly form the basis for ensemble perception tasks (Figure 6A, though see (Balas, 2016) for questions of whether our particular candidate encoding can quantitatively predict judgments of numerosity). In contrast, a handful of ensemble statistics cannot obviously support rich scene perception, and without specifying the statistics it is not even clear that they can support the rich perception of ensembles. As Huan et al. (2017) point out, referring to an array of letters (Figure 6B), observers likely know quite a bit about ensembles:

> "Is that really all they see, [3-4 items] perhaps augmented by some summary statistics? A moment's reflection indicates that, if only they were asked, subjects could report much more – one certainly perceives that there are many black

> marks, that they are arranged in rows and columns, in a rectangular array,… against a bright homogeneous background… [these percepts are] typically taken for granted rather than included in the catalog of conscious contents… While subjects may not be able to recognize specific identities,… they can effortlessly report that what they saw were letter-like figures."

A critical point here, however, is that while "some" unspecified summary statistics cannot obviously predict this rich percept, a set of many summary statistics can. As the syntheses in Figure 6A show, the proposed encoding clearly preserves sufficient information to answer questions about the distribution of line orientations, including the mean and variance. In addition, it preserves enough information to tell that the stimulus is made up of black lines on a light background, an important characteristic most likely available to the

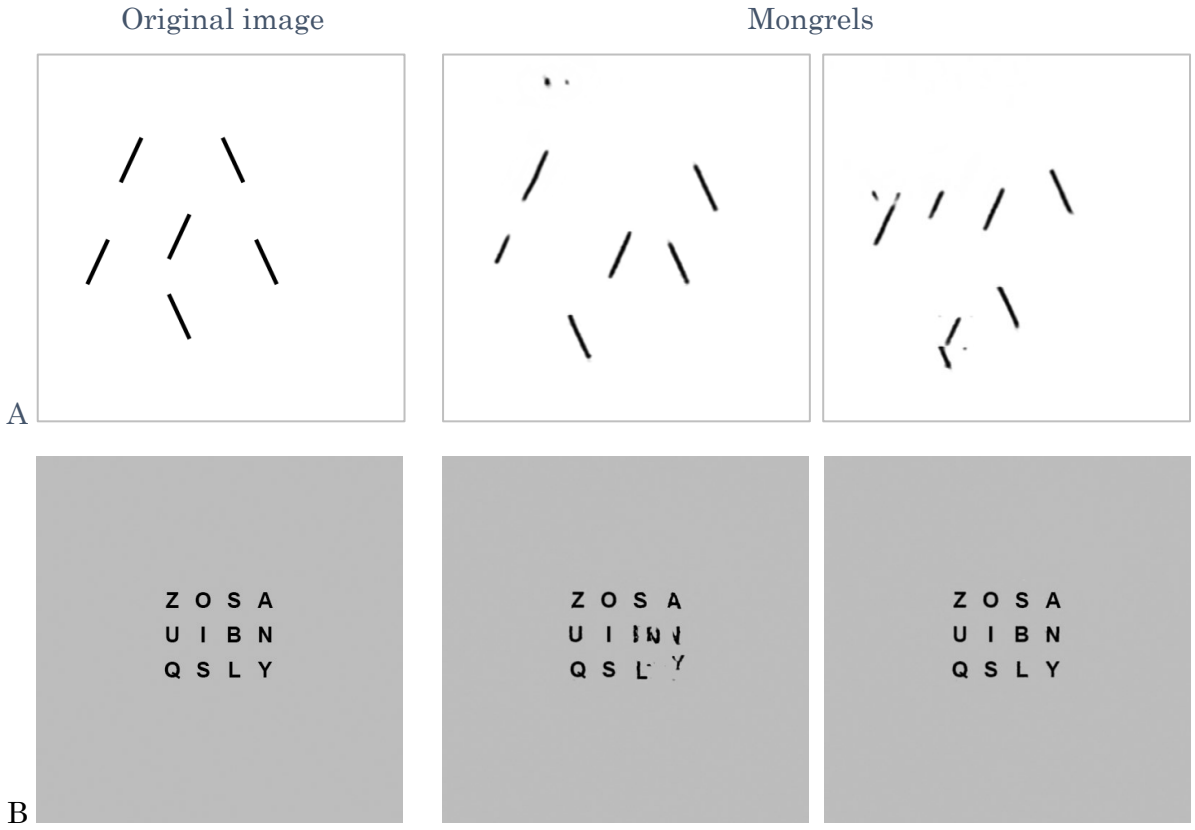

Figure 6. The proposed set of summary image statistics encode considerable information about sets of similar items. A. Original set of oriented lines (left), and two syntheses visualizing the information available (right). Modeled with the fixation 10 degrees to the right of the central target, where each line is 1 degree in length. B. Array of letters (left) like that in Sperling (1960). Syntheses (right) predict that peripheral vision can discern the structure and appearance of the array, and even support identifying the majority of the letters. Fixation on the "I", modeled as subtending approximately 2 degrees of visual angle.

observer. The sizes and orientations of items are also largely preserved, but location information is lost; the lost information perhaps partially explains the difficulty reporting the features of a *particular item*, e.g. in (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; Fischer & Whitney, 2011). The mongrels of the letter arrays (Figure 6B) similarly indicate that the encoding preserves precisely the sort of information enumerated by Huan et al. In addition, it appears that sufficient information survives to recognize 10-12 of the letters – far greater than the average 4.3 items available for immediate report, but comparable to the 9.1 letters estimated to be available by partial report (Sperling, 1960).

Perhaps previous theories have described the representation of ensemble statistics instead of image statistics as merely a rhetorical figure of speech. It is easier to get intuitions about and to enumerate the mean size and orientation of a set of items than to think about more abstract image statistics. In addition, researchers may have inadvertently implied that their theories required only a few statistics because of the difficulty coming up with a long list of plausible ones. Both points, however – image statistics, and lots of them – are critical to the argument that such an encoding could underlie the richness of perception. It is important to be explicit. When Cohen et al. (2016) refer to a "single summary statistic", this could in principle refer to a single high-dimensional vector – which is, after all, what underlies their demos from Freeman and Simoncelli (2011) and Oliva and Torralba (2006) – but if so they risk confusing their readers.

Returning to the question posed in Section 2.1: Does it matter, when asking whether vision is impoverished, if tasks like search and change blindness are difficult because of the limits of attention or the limits of peripheral vision? Clearly it does. The attention explanation requires an additional mechanism to explain why observers easily get the gist of scenes and sets, whereas the peripheral vision explanation does not. Furthermore, attentional capture theory was developed to make sense of the apparent chicken-and-egg

Figure 7. Looking for one's mug on the desk, the same peripheral encoding that predicts difficult search and change blindness provides ample information to locate the desk, notice salient objects, and guide eye movements to gather additional information. Fixation as indicated by the red cross.

problem of how one can successfully direct attention if perception without attention is so poor. The peripheral vision explanation has no such problem. Consider looking for one's mug in the office scene in Figure 7. Starting with a central fixation, the proposed encoding scheme provides ample information for locating the desk, noticing the salient painting on the wall to the left, and noticing the person seated at the desk. That glance may not, however, preserve enough information to immediately find one's mug. One cannot recover the information lost in peripheral vision without an eye movement, but the information that remains is capable of supporting performance of many tasks, from guiding eye movements, through some object recognition tasks, to getting the gist of a scene and navigating the world.

3. A proposal for an additional capacity limit: Limited decision complexity

3.1 Other difficult tasks may be inherently difficult

Given the strengths of peripheral vision, it is not surprising that observers can easily get the gist of a scene or set. The limitations of peripheral vision can explain many of the phenomena previously taken as evidence that perception is poor without attention. This

paper began, however, by also enumerating a second set of phenomena that suggest that attention itself is limited, as is visual working memory (Figure 1). These phenomena clearly cannot be explained by peripheral vision alone. Peripheral vision could be a *factor* – inattentional blindness, multiple object tracking (MOT), and visual working memory (VWM) tasks often utilize crowded displays, and typical dual-task experiments assign one task to peripheral vision. However, a number of inattentional blindness (Mack & Rock, 1998; Levin & Simons, 1997) studies have forced fixation, and found that knowing the task matters. Visual working memory studies (e.g. (Tamber-Rosenau, Fintzi, & Marois, 2015; Adam, Vogel, & Awh, 2017) have controlled for peripheral crowding and found similar memory limits. Typical dual-task experiments (e.g. (VanRullen, Reddy, & Koch, 2004)) hold fixation and the display constant, and vary the number of tasks; though peripheral discriminability does appear to be a factor in dual-task difficulty (Rosenholtz, Huang, & Ehinger, 2012), it cannot explain why many dual-tasks are more difficult than their component single tasks. Other tasks may also encounter additional limits; search and change detection, for instance, may be more difficult than predicted from peripheral vision alone (Rosenholtz, 2017). There must be some other capacity limit(s).

It may be tempting, at least in the case of dual-task performance, inattentional blindness, and MOT, to fall back on selective attention theory to explain these results. Quite a bit of the evidence for that theory, however, had a peripheral vision confound, and peripheral vision offers a more parsimonious account. At minimum, it would seem a useful exercise to start from scratch in examining the remaining capacity limit(s). For a detailed argument for why we need to look for a different sort of capacity limit, and for different mechanisms for dealing with that limit, see (Rosenholtz, 2017).

I argue that this second group of tasks is inherently difficult. Consider a typical VWM task. An observer is shown an array of k items, such as colored disks (Figure 1, lower

right). After a delay, the experimenter then presents another array. This array either duplicates the original, or differs in the color of one of the *k* disks. In the traditional way of thinking of this task, the observer has *n* memory slots to fill with features from each of the *k* disks. By examining how performance varies as a function of the number of items in the display, one can supposedly infer the number of slots. Based on this logic, researchers have concluded that a typical observer has only around 4 slots, suggesting a very limited capacity for VWM (Luck & Vogel, 1997).

This logic, however, makes strong assumptions about the mechanisms underlying VWM: a fairly strong "brain as computer" analogy in which memory works by putting information into some kind of storage for later retrieval. Let us think of the VWM task at a more basic level, without such assumptions. The observer must discriminate between the array to be remembered and all other similar arrays in which one item differs. (The VWM paradigm sometimes instead asks the observer to specify the features of a particular post-cued item. For the sake of argument, I assume that changing the task in that way does not fundamentally change its inherent difficulty.) One could imagine that, for a randomly chosen initial display, this discrimination would require a fairly complex classifier. Just how complex would depend upon the feature space in which the classifier operates. The brain's feature space seems unlikely to be a vector of *k* colors. Put another way, arrays of colored disks likely occupy a very small region within the brain's "perceptual encoding space" – the space of images one is likely to see, represented in whatever high-dimensional encoding the visual system employs. Discriminating between such similar images might be quite difficult. A very similar story applies to tasks such as reporting a post-cued member of an ensemble. This task is essentially a VWM task, and is likely hard at least in part for the same reason.

One can make a related argument that MOT tasks (Figure 1, upper left) are inherently difficult. In the traditional interpretation, the visual system has *m* attentional spotlights to deploy. Based on performance one can infer *m*, and as *m* is low, one concludes that attention has limited capacity (Pylyshyn & Storm, 1988). However, as with VWM, this account makes strong assumptions about the mechanisms involved. At a basic level, if the observer must track *k* of *n* items, then on each frame they must distinguish the actual *k* targets from *n*-choose-*k* other possible combinations of *k* items. In the case of tracking 4 of 9 items, for instance, the observer must distinguish the actual 4 targets from 125 other possibilities! One might again imagine that this task is inherently difficult in the abstract, though of course motion cues make the task more tractable.

Consider also typical dual-task experiments (e.g. (VanRullen, Reddy, & Koch, 2004); Figure 1, upper right). The observer is asked either to complete a single peripheral task, or to perform that task as well as a central task. For instance, the observer might specify

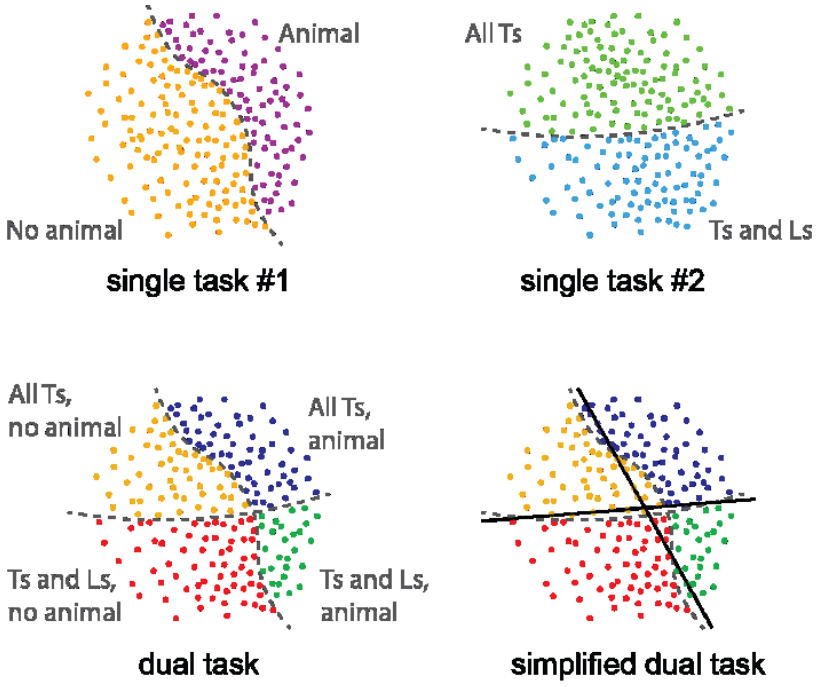

Figure 8. Dual tasks are inherently more complex than their component single tasks. Here, two 2AFC tasks (top) become a 4AFC dual task (bottom left). If there exists a limit on task complexity, the observer will have to simplify this task (bottom right, solid lines), making errors.

whether a peripheral cube is upright or inverted, while also indicating whether a central array contains all the same letter (all Ls or all Ts) or different letters (both Ls and Ts). Both the central and peripheral task involve distinguishing between two alternatives. The dual task involves distinguishing between four possibilities (Figure 8). This renders dual tasks inherently more complex, and this complexity may be why multitasking is often difficult.

Looked at in this way, these tasks appear to be inherently difficult almost regardless of the underlying representation (Tsotsos, 1990). Inherently difficult decisions suggest the involvement of late, decision-level mechanisms, and decision-level limits. The exact nature of these decision-level limits remains unclear. It cannot simply be a limit on task difficulty. Dual-task experiments controlled for difficulty of the component tasks (e.g. (VanRullen, Reddy, & Koch, 2004)); if task difficulty were the only issue, all dual-tasks would be equally hard (Rosenholtz, 2017).

Based on the arguments above, perhaps decision mechanisms instead face a limit on task *complexity* (Rosenholtz, 2017). This could take a number of different forms. Our cognitive processes might be limited in the number of dimensions (or neurons) one could use to make a decision; in the number of linear hyperplanes out of which one could form a decision boundary; or in the curviness of that boundary, etc. Such a complexity limit might exist for the usual reasons, e.g. limits on the size of the brain (Tsotsos, 1990). In addition, in learning to perform a classification task, limiting decision complexity might be a way to avoid overfitting the decision boundary.

3.2 Limited decision complexity: Implications for a rich subjective impression and real world vision

Let us consider a couple of examples, both to get used to thinking about decision complexity, and to tie this proposal back to the awareness puzzle and the success of real-

world vision. The reader could, at this point, have an important question: I have argued that VWM is limited because it is an inherently complex task; how complex, then, is scene perception? In both cases one might think of the implicit task as distinguishing between seen and not seen – essentially as localization in some perceptual encoding space (Figure 9). In the proposed theory, what the observer *knows* about the stimulus as a result of performing this localization task – what they *perceive* – is determined by the classification into seen and not seen. If the classification boundary confuses two images then *from this classification task alone* (a point we will consider shortly), the observer cannot perceive the differences between them. Lower precision at this task might require less effort, but at the cost of confusing more unseen stimuli with the one actually seen; with lower precision, the

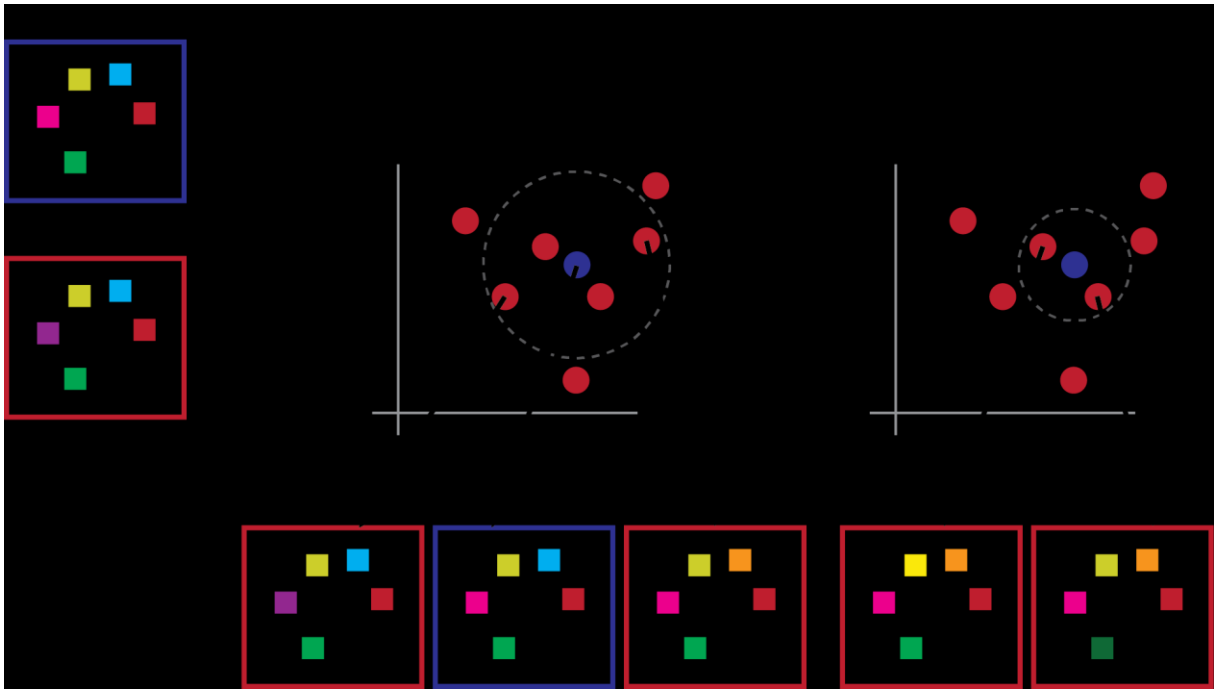

Figure 9. At a basic level, we can think of Visual Working Memory (VWM) tasks as distinguishing between the observed stimulus and all similar stimuli that differ in one of the items (upper left). If we think of each stimulus image as represented by a high-dimensional vector in some perceptual encoding space (shown here with only two dimensions for simplicity), then we can think of this discrimination as a classification. Dashed lines indicate two possible classification boundaries. The boundary on the right is more precise, distinguishing the observed array (blue) from most other arrays, except those with small color differences. Capacity limits may prohibit such a precise classification, perhaps because they limit complexity, e.g. curvature of the decision boundary. Instead, the brain may be forced to use a less precise decision boundary, such as that shown on the left. This may require less effort, but leads to more significant confusions between the seen and unseen arrays.

observer knows less. With more effort, the observer might be able to utilize a more complex – higher curvature – classification boundary between seen and unseen stimuli, making fewer errors. However, if there exists a limit on decision complexity, that means that precision and knowledge about the stimulus are limited.

When we speak of a *limit*, this implies the existence of a single cap that all visual tasks must obey. Here I have been assuming that VWM tasks encounter this limit, making it appear that we can remember only about four items at a time. If our scene perception encounters the same limit, how rich should we expect that percept to be? The answer depends fundamentally on the underlying perceptual encoding, which remains essentially unknown. However, we can get a hint of the answer from the following mini-experiment:

Let us take our candidate perceptual encoding from a convolutional neural network (CNN), known as VGG-16, which was trained to perform invariant object recognition in real-world scenes (Simonyan & Zisserman, 2014).[1] CNNs have recently become very popular, as for the first time they allow computer vision to approach human performance on certain proscribed visual tasks. Researchers have also shown certain similarities between the representations learned by CNNs and those found in monkey physiology (Yamins, et al., 2014). We took a set of arrays of 8 colored squares against a gray background, and fed them into the network to generate a feature vector for each image. For the feature vector, we used the last representational layer (the "last fully connected layer") of the network; it is common in computer vision to use this layer as the input to classifiers.

These images are confusable in a standard VWM task; we can measure the distance between their VGG-16 feature vectors to give us an estimate of the available precision for

---

[1] Note that this encoding is not foveated. Despite the importance of peripheral vision for understanding many relevant perceptual phenomena, for this mini-experiment we use an encoding that does not depend upon distance from the point of gaze.

localizing any image in perceptual encoding space. Given that same uncertainty, how well could we instead pinpoint a natural scene? We took a set of similar street scenes, computed their VGG-16 feature vectors, and then asked what scenes would be difficult to discriminate, given the same precision inferred from the VWM stimuli. The top left of

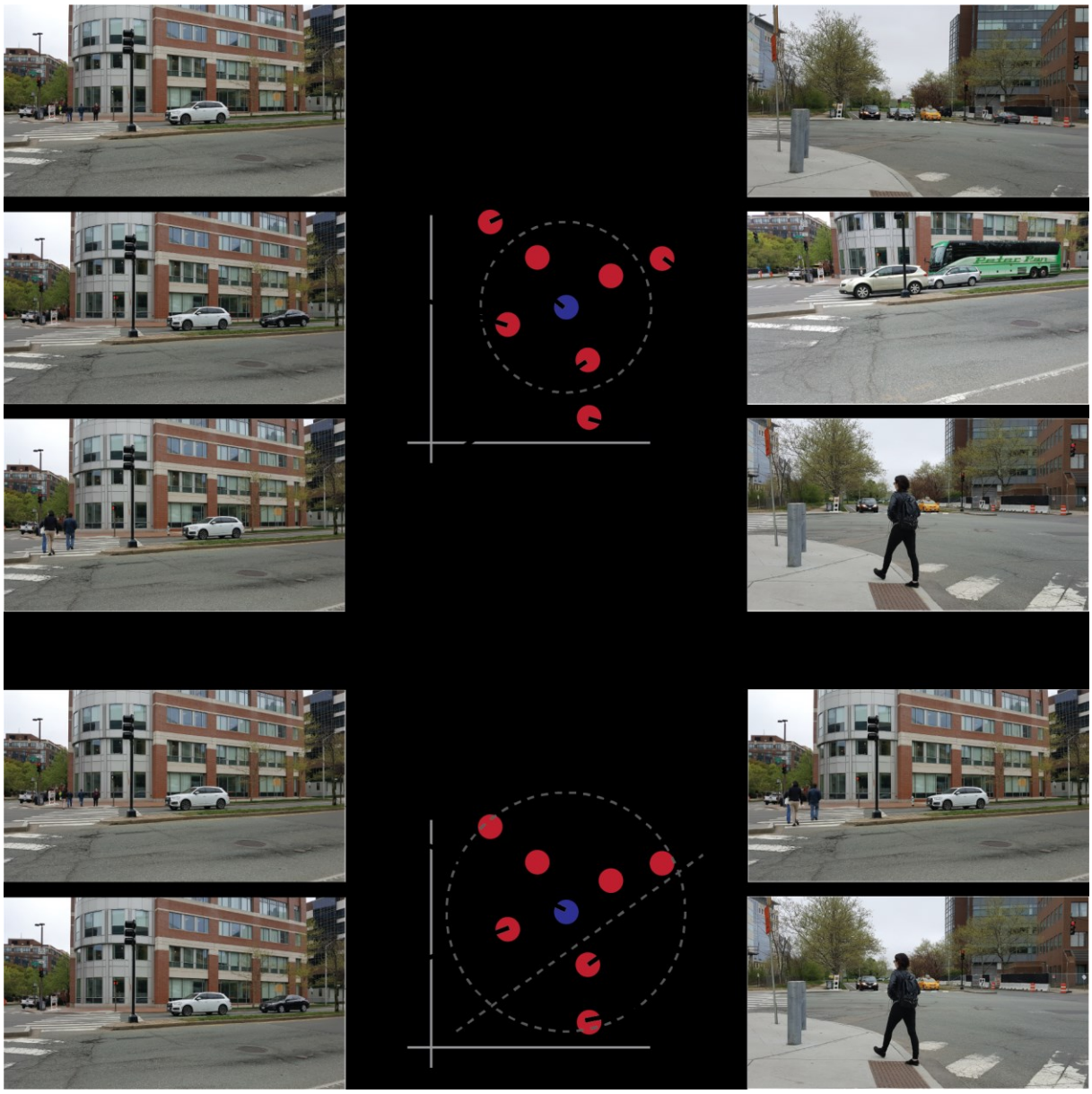

Figure 10. (top) The three confusable images on the left have similar mean discriminability as arrays of 8 colored squares, given the perceptual encoding space of the VGG-16 neural network. The three images on the right are less confusable with these images, according to discriminability in that feature space. (bottom) Switching to a different task can lead to new understanding of the scene. At the next moment, the visual system might attempt to discriminate scenes with nearby pedestrians (right) from those for which the pedestrians were absent or farther away (left).

Figure 10 shows a set of three confusable scenes, according to this metric. However, by this metric these scenes are discriminable from those in the top right.

The first thing to note is that a distance metric applied to the last fully connected layer of VGG-16 seems to give us a reasonable measure of perceptual similarity. It is difficult to distinguish arrays of randomly-colored squares from each other (Figure 9), and analogously difficult to distinguishing the confusable scenes in Figure 10. Those scenes do differ: the camera angle has changed somewhat, and the location and number of vehicles and pedestrians has changed. The less confusable scenes in the top right appear more readily discriminable. So the mini-experiment is a good first attempt. More importantly, note that for the same amount of uncertainty that makes an 8-item VWM task hard, one can pinpoint a scene fairly well. No doubt the visual system has developed to make this so. In a plausible perceptual encoding space, the same precision can specify either "an array of about 8 items of random color and position," or mostly determine the scene, plus or minus some small changes. This suggests there is real hope for a unified explanation. The same inference limits that make VWM difficult allow a rich subjective experience of the real world.

In real-world vision, we often need to know more about the scene; for example when driving we must estimate the 3D location of the pedestrians in order to judge whether we can turn left. Thankfully our perception is not merely limited by the results of performing the "gist" task just described. In the next instant, the observer can perform a different task, i.e. pose another question and make a new inference. In this case, the observer might next ask about the location of the pedestrians, i.e. classify the scene into those containing near vs. far pedestrians. The layout information gained from the "gist" inference could help refine this task by indicating where pedestrians are likely to be. The pedestrian localization task, because it does not require detailed knowledge of the rest of the scene, could be less

complex. Even if, in our demo, near and far pedestrians could not be discriminated in the "gist" task because of complexity limits, they might nonetheless be discriminated in a pedestrian-localization task. The observer gains additional understanding about the pedestrians at the expense of comprehension of the scene as a whole. Many typical real-world tasks probably have low complexity relative to the limit – again, the brain has likely developed its representation to make this the case. As a result, while estimating the 3D position of the pedestrian the observer may not completely lose the gist of the scene as a whole, but may just become more imprecise at localizing the scene in the perceptual encoding space.

Similarly, in the VWM task, the lack of precision when trying to remember the entire array does not imply that the observer cannot discriminate whether a particular square is red or blue. If that is the task, for instance if one of the squares is pre-cued ("remember this one"), then the observer can set up a relatively simple classifier to discriminate the color of that square, again likely at the expense of some details about the set as a whole.

In this scheme, then, real-world vision consists of a series of tasks – questions the observer asks him- or herself, perhaps unconsciously (Helmholtz, 1867) – and perception is composed of the outcomes of those inferences. If a given task exceeds the complexity limit, the observer will need to perform a simpler task instead, perhaps without awareness that they are doing so.

This notion of switching between tasks, each leading to different understanding of the scene, may sound a bit like the proposal from Treisman (2006) that the observer can switch between different attentional "modes". Treisman suggested that attention is a limited resource with some flexibility in how diffusely it can be allocated. Attending to a scene or a set yields holistic properties without the details, whereas object-based attention

yields understanding of the object at the expense of the scene. (Other researchers have made related proposals, e.g. (Nakayama, 1990; Van Essen, Olshausen, Anderson, & Gallant, 1991; Rensink, Change blindness: Implications for the nature of attention, 2001). It requires little effort to see relationships between switching tasks because of limited decision capacity and switching mode because of limited attention. Treisman's proposal of additional attentional modes appeared to point towards a solution to the problems with earlier versions of selective attention theory. It paved the way to further studies on what information becomes available upon diffusely attending to a scene or a set (Alvarez, 2011; Fei-Fei, Iyer, Koch, & Perona, 2007; Greene & Oliva, 2009; Leib, Kosovicheva, & Whitney, 2016). However, this proposal also raised a number of questions. What, for instance, are the mechanisms associated with diffuse attention? How does the brain switch attentional modes, and how do upstream processes deal with potentially dramatic changes in the encoding of available information? How many different modes of attention are there? How can we characterize, and thus predict, the limited detail available under diffuse attention to a scene? If diffuse attention and focal attention are two different mechanisms for dealing with a single capacity limit, then how should we conceptualize that capacity limit? Even if one thinks of the present proposal as a mere reframing of different attentional modes in terms of switching tasks to deal with limited decision complexity, this reframing illuminates a new path forward in understanding the mechanisms, their flexibility, and the nature of the limit.

For instance, different attentional modes suggest that from moment to moment the information encoded by the visual system can change dramatically with the focus and type of attention. Later processes must somehow deal with the highly dynamic nature of the encoded information. Changing task to accommodate limited decision complexity does not raise the same issues. Rather, each new task requires a late mechanism to set up a new

classifier and interprets its results (though one may perhaps see effects of this mechanism early in visual processing as well). This theory presumes that, to a first approximation, changing the task changes neither the encoding nor the available information. Rather, each new query changes what we *know*. The answer to the question of whether the pedestrians are near or far gives us new understanding of the scene.

Similarly, it has been unclear what capacity limit might be satisfied both by focal attention to an object and by diffuse attention to a scene; i.e. in what sense these two attentional modes might be equivalent in terms of use of available resources. Several researchers have speculated about the answer to this question (Nakayama, 1990; Van Essen, Olshausen, Anderson, & Gallant, 1991). Van Essen et al., for instance, suggested that the visual system might always have access to an approximately 25 × 25 array of feature vectors. These feature vectors could be spread either over an object or over the entire scene, and might derive from any layer in the visual processing hierarchy. While these proposals are intriguing, it has not been obvious how to advance these theories, or what alternatives might exist. On the other hand, while the exact nature of the decision complexity limit remains unclear, there would appear to be a viable path forward. We have considerable understanding of human behavioral limits, and can use those limits to look for a consistent complexity limit such as those described above: number of hyperplanes, number of dimensions, curvature of the decision boundary and so on. Machine learning also has a concept of decision complexity, and can provide other forms that this limit might take, e.g. (Vapnik & Chervonenkis, 1971). Of course, looking for a consistent limit requires a model of the perceptual encoding space, but vision research has advanced to the point where one may feasibly use either computational models, such as trained CNNs, or rich, high-dimensional data from physiology, e.g. from fMRI. Understanding of possible decision limits, in turn, should make testable predictions of what tasks observers can and cannot do.

Unlike the previous proposal that attention is limited to either a focal mode that provides object properties, or a diffuse mode that leads to scene and set properties, one would expect that a diverse set of decision boundaries satisfy the complexity limit. Researchers have suggested only a handful of mechanisms for dealing with attentional limits: e.g. attend to only this object, only this color, or only this location. One would expect that mechanisms for dealing with limited complexity might be considerably more diverse. If the ideal task required too wiggly a decision boundary, the observer would have to perform a simpler task instead, and would make errors as a result (Figure 8, bottom right). Simplifying strategies might include setting up a classifier to identify only one object, only objects with a certain color, or only the object at a particular location, in an obvious parallel with attentional theories. However, the visual system may have available more general strategies for "cutting corners" – literally (Figure 8) – in order to simplify an overly complex decision boundary.

4. Conclusions: A proposed unifying explanation

I have argued that the strengths and limitations of visual perception result from constraints arising from both perceptual encoding and decision processes. A visual task can be difficult because of limits in either of these processes.

First, a striking number of puzzling visual phenomena can be explained simply by the information preserved and lost in peripheral vision. Peripheral vision appears to encode its inputs in terms of a rich set of summary image statistics, computed by pooling image measurements across sizeable regions of the visual field. These regions grow – and the resulting summary statistics become increasingly less informative – with distance from the point of gaze. At a given moment, the current fixation largely determines the information available across the field of view. If information needed for a task does not survive the

peripheral encoding, that task will be difficult. To gather more information, observers must move their eyes. Losses in this encoding lead to poor performance on a number of visual tasks (difficult search, change blindness), while preserving sufficient information to make other tasks relatively easy (easy search, and getting the gist of a scene or set), and to support our rich percept of the world.

However, some tasks are difficult even if the necessary information survives both peripheral vision and the perceptual encoding stages more generally. I have argued that the second big piece of the solution has to do with decision limits, and in particular, limits on decision complexity. Dual tasks may be more difficult than single tasks because they are inherently more complex. Inattentional blindness – the inability to perform a task when it is unexpected – may occur when limits on decision complexity preclude performing both the nominal task and, by chance, also the unexpected task. MOT and VWM may both be inherently complex tasks, leading to apparent limits on the number of items that can be tracked or remembered.

If an additional capacity limit applies late in processing, at the decision stage, then this raises the intriguing possibility that it might be a general-purpose cognitive capacity limit, rather than a limit solely on visual processing. In fact, there exists some evidence for this, from analysis of individual differences. Huang et al. (2012) found correlated performance at a wide range of tasks, including search, counting, tracking, response selection, short-term memory, visual marking, task switching, and mental rotation.

That perception results from inference suggests that there is some truth to the "illusion" theories of awareness. One perceives the results of inference, not some image captured by the eye-as-camera, and projected onto an internal screen for viewing by the homunculus. In this sense, perception is inherently something of an illusion. However, the illusion is not as extreme as previously thought, because vision is less impoverished, and

thus the rich percept less surprising. Rather, tasks that seem to show impoverished vision may simply be difficult tasks, either due to the encoding or due to limits on inference processes. On the other hand, perception is rich, and real-world vision successful, because the information for many tasks survives encoding losses, and that encoding evolved to make those tasks relatively simple. I have argued that many phenomena – search, set perception, scene perception, visual working memory, multiple object tracking, dual-task, and change blindness – may encounter the same limits on both the information encoded and the complexity of decisions. Given those limits, some tasks may simply be inherently difficult, and others easy. If so, there is no need to ponder why, for instance, we get a rich subjective impression and yet do poorly at certain tasks; no need to postulate that the details are puzzlingly inaccessible for decision and action. If a unified explanation is possible, there is no awareness puzzle.

Acknowledgements

The work described here was funded in part by NIH-NEI EY021473 and NIH NEI R21-EY019366 to Ruth Rosenholtz, and NSF/BMBF IIS-1607486 to Ruth Rosenholtz and Christoph Zetzsche. Thanks to Shaiyan Keshvari and Yrvine Thelusma for help with the VGG-16 experiment, to Benjamin Wolfe, Shaiyan Keshvari, Michael Cohen, and Dian Yu for useful discussions, and to Todd Horowitz for drawing the connection between decision-level mechanisms and evidence of a general-purpose cognitive capacity limit.

References

Adam, K. C., Vogel, E. K., & Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology, 97*, 79-97.

Alvarez, G. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15*, 122-131.

Andriessen, J. J., & Bouma, H. (1976). Eccentric vision: Adverse interactions between line segments. *Vision Research, 16*(1), 71-78.

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

Balas, B. J. (2016). Seeing number using texture: How summary statistics account for reductions in perceived numerosity in the visual periphery. *Attention, Perception, & Psychophysics, 78*(8), 2313-2319.

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

Blackmore, S. J., Brelstaff, G., Nelson, K., & Troscianko, T. (1995). Is the richness of our visual world an illusion? Transsaccadic memory for complex scenes. *Perception, 24*, 1075-1081.

Block, N. (2011). Perceptual consciousness overflows cognitive access. *Trends in Cognitive Sciences, 15*(12), 567-575.

Carrasco, M., & Frieder, K. S. (1997). Cortical magnification neutralizes the eccentricity effect in visual search. *Vision Research, 37*, 63-82.

Carrasco, M., & Yeshurun, Y. (1998). The contribution of covert attention to the set-size and eccentricity effects in visual search. *J. of Exp. Psychology: Human Perception & Performance, 24*, 673-692.

Carrasco, M., Evert, D. L., Chang, I., & Katz, S. M. (1995). The eccentricity effect: Target eccentricity affects performance on conjunction searches. *Perception & Psychophysics, 57*, 1241-1261.

Carrasco, M., McLean, T. L., Katz, S. M., & Frieder, K. S. (1998). Feature asymmetries in visual search: Effects of display duration, target eccentricity, orientation, & spatial frequency. *Vision Research, 38*, 347-374.

Chang, H., & Rosenholtz, R. (2016). Search performance is better predicted by tileability than by the presence of a unique basic feature. *Journal of Vision, 16*(10), 13.

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

Chong, S.-C., & Treisman, A. (2005). Attentional spread in the statistical processing of visual displays. *Percept. & Psychophys., 66*, 1282-1294.

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the bandwidth of perceptual experience. *Trends in Cognitive Sciences, 20*(5), 324-335.

Dennett, D. C. (1991). *Consciousness explained.* Boston: Little Brown.

Dennett, D. C. (1998). No bridge over the stream of consciousness. *Behavioral and Brain Sciences, 21*, 753-754.

Ehinger, K. A., & Rosenholtz, R. (2016). A general account of peripheral encoding also predicts scene perception performance. *Journal of Vision, 16*(2), 13.

Erkelens, C. J., & Hooge, I. T. (1996). The role of peripheral vision in visual search. *J. of Videology, 1*, 1-8.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision, 7(1):10*, 1-29.

Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. Journal of Neurophysiology. *Journal of Neurophysiology, 106*(3), 1389-1398.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Perception & Performance, 18*(4), 1030-1044.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience, 14*(9), 1195-1201.

Geisler, W. S., Perry, J. S., & Najemnik, J. (2006). Visual search: The role of peripheral information measured using gaze-contingent displays. *Journal of Vision, 6*(9), 858-873.

Gheri, C., Morgan, M. J., & Solomon, J. A. (2007). The relationship between search efficiency and crowding. *Perception, 36*, 1779-1787.

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cogn. Psychol., 58*, 137-176.

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception & Performance, 35*(3), 718-734.

Haberman, J., & Whitney, D. (2011). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In J. Wolfe, & L. Robertson (Eds.), *A Festschrift in honor of Anne Treisman.*

Helmholtz, H. v. (1867). *Handbuch der Physiologischen Optik.* Leipzig: Voss.

Henderson, J. M., & Hollingworth, A. (1999). The role of fixation position in detecting scene changes across saccades. *Psychological Science, 10*, 438-443.

Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *J. of Exp. Psych.: Human Perception & Performance, 28*(1), 113-136.

Huan, A. M., Tononi, G., Koch, C., & Tsuchiya, N. (2017). Are we underestimating the richness of visual experience? *Neuroscience of Consciousness*, 1-4.

Huang, L., Mo, L., & Li, Y. (2012). Measuring the interrelations among multiple paradigms of visual attention: An individual differences approach. *Journal of Experimental Psychology: Human Perception & Performance, 38*(2), 414-428.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience, 2*(3), 194-203.

Keshvari, S., & Rosenholtz, R. (2016). Pooling of continuous feature provides a unifying account of crowding. *Journal of Vision, 16*(3), 39.

Koch, C., & Crick, F. (2001). The zombie within. *Nature, 411*, 893.

Kooi, F. L., Toet, A., Tripathy, S. P., & Levi, D. M. (1994). The effect of similarity and duration on spatial interaction in peripheral vision. *Spatial Vision, 8*(2), 255-279.

Lamme, V. (2010). How neuroscience will change our view on consciousness. *Cognitive Neuroscience, 1*, 204-220.

Leib, A., Kosovicheva, A., & Whitney, D. (2016). Fast ensemble representations for abstract visual impressions. *Nature Communications, 7*, 13186.

Lettvin, J. Y. (1976). On seeing sidelong. *The Sciences, 16*(4), 10-20.

Levin, D. T., & Simons, D. J. (1997). Failure to detect changes to attended objects in motion pictures. *Psychonomic Bulletin & Review, 4*, 501-506.

Livne, T., & Sagi, D. (2007). Configuration influence on crowding. *Journal of Vision, 7*(2), 4, 1-12. doi:10.1167/7.2.4

Loftus, G. R., & Ginn, M. (1984). Perceptual and conceptual masking of pictures. *J. Exp. Psychol. Learn. Mem. Cogn., 10*(3), 435-441.

Loschky, L. C., Sethi, A., Simons, D. J., Pydimarri, T. N., Ochs, D., & Corbeille, J. L. (2007). The importance of information localization in scene gist recognition. *J. Exp. Psych.: Human Perception & Performance, 33*(6), 1431-1450.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*, 279-281.

Mack, A., & Rock, I. (1998). *Inattentional blindness.* Cambridge, MA: MIT Press.

Manassi, M., Sayim, B., & Herzog, M. H. (2012). Grouping, pooling, and when bigger is better in visual crowding. *Journal of Vision, 12*(10), 13. doi:10.1167/12.10.13

Martelli, M., Majaj, N. J., & Pelli, D. G. (2005). Are faces processed like words? A diagnostic test for recognition by parts. *Journal of Vision, 5*(1), 6.

Nakayama, K. (1990). The iconic bottleneck and the tenuous link between early visual processing and perception. In C. Blakemore (Ed.), *Vision: Coding and Efficiency* (pp. 411-422). Cambridge University Press.

Noë, A. (2002). Is the visual world a grant illusion? *Journal of Consciousness Studies, 9*, 1-12.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Prog. Brain Res., 155*, 23-36.

O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian J. of Psychology/REvue Canadienne de Psychologie, 46*(3), 461-488.

O'Regan, J. K., Deubel, H., Clark, J. J., & Rensink, R. A. (2000). Picture changes during blinks: Looking without seeing and seeing without looking. *Visual Cognition, 7*(1-3), 191-211.

O'Regan, J. K., Rensink, R. A., & Clark, J. J. (1999). Change-blindness as a result of 'mudsplashes'. *Nature, 398*(4), 34.

Parker, R. E. (1978). Picture processing during recognition. *J. Exp. Psych.: Human Perception & Performance, 4*(2), 284-293.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, J. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*, 739-744.

Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience, 11*, 1129-1135.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis., 40*(1), 49-71.

Potter, M. C. (1975). Meaning in visual search. *Science, 187*, 965-966.

Potter, M. C., & Fox, L. F. (2009). Detecting and remembering simultaneous pictures in a rapid serial visual presentation. *J. Exp. Psych.: Human Perception & Performance, 35*, 28-38.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision, 3*(3), 1-19.

Rensink, R. A. (2001). Change blindness: Implications for the nature of attention. In M. R. Jenkin, & L. R. Harris (Eds.), *Vision and Attention* (pp. 169-188). New York: Springer.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8*, 368-373.

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Research, 39*, 3157-3163.

Rosenholtz, R. (2011). What your visual system sees where you are not looking. *Proc. SPIE 7865, Hum. Vis. Electron. Imaging, XVI.* San Francisco.

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual Rev. of Vision Sci., 2*(1), 437-457.

Rosenholtz, R. (2017). Capacity limits and how the visual system copes with them. *Journal of Imaging Science and Technology (Proc. HVEI, 2017)*, 8-23.

Rosenholtz, R., Huang, J., & Ehinger, K. A. (2012). Rethinking the role of top-down attention in vision: effects attributable to a lossy representation in peripheral vision. *Frontiers in Psychology, 3:13*. doi:doi: 10.3389/fpsyg.2012.00013

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(4):14*, 1-17.

Rosenholtz, R., Sharan, L., & Park, E. (under revision). Inattentional blindness: Interaction of gaze pattern and task demand. *Attention, Perception, and Psychophysics*.

Rousselet, G. A., Joubert, O., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition, 12*(6), 852-877.

Sayim, B., Westheimer, G., & Herzog, M. H. (2010). Gestalt factors modulate basic spatial vision. *Psychological Science, 21*(5), 641-644.

Scholl, B. J. (2000). Attenuated change blindness for exogenously attended items in a flicker paradigm. *Visual Cognition, 7*(1-3), 377-396. doi:10.1080/135062800394856

Scott-Brown, K. C., Baker, M. R., & Orbach, H. (2000). Comparison blindness. *Visual Cognition, 7*, 253-267.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28*, 1059-1074.

Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition.* arXiv Technical Report.

Smith, M. E., Sharan, L., Park, E., Loschky, L. C., & Rosenholtz, R. (under revision). Difficulty detecting changes in complex scenes depends in part upon the strengths and limitations of peripheral vision. *Journal of Vision*.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied, 74*(11), 1-29.

Tamber-Rosenau, B. J., Fintzi, A. R., & Marois, R. (2015). Crowding in visual working memory reveals its spatial resolution and the nature of its representations. *Psychological Science, 26*(9), 1511-1521.

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics, 51*(6), 599-606.

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience, 17*, 450-461.

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

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol., 12*, 97-136.

Tsotsos, J. K. (1990). Analyzing vision at the complexity level. *Behavioral and Brain Sciences, 13*, 423-469.

Van Essen, D. C., Olshausen, B., Anderson, C. H., & Gallant, J. L. (1991). Pattern recognition, attention, and information bottlenecks in the primate visual system. *Proc. SPIE 1473, Visual Information Processing: From Neurons to Chips*, (pp. 17-28).

VanRullen, R., Reddy, L., & Koch, C. (2004). Visual search and dual tasks reveal two distinct attentional resources. *J. Cogn. Neurosci., 16*, 4-14.

Vapnik, V. N., & Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*.

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

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided Search: An alternative to the Feature Integration Model for Visual Search. *Journal of Experimental Psychology: Human Perception & Performance, 15*(3), 419-433.

Wolfe, J. M., Vo, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and non-selective pathways. *Trends in Cognitive Sciences, 15*(2), 77-84.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS, 111*, 8619-8624. doi:10.1073/pnas.1403112111

Zelinsky, G. J. (2001). Eye movements during change detection: Implications for search constraints, memory limitations, and scanning strategies. *Perception & Psychophysics, 63*(2), 209-225.

Zhang, X., Huang, J., Yigit-Elliot, S., & Rosenholtz, R. (2015). Cube search, revisited. *Journal of Vision, 15*(3), 9.