https://doi.org/10.2352/ISSN.2470-1173.2017.14.HVEI-111

Capacity limits and how the visual system copes with them

Ruth Rosenholtz; Massachusetts Institute of Technology, Department of Brain & Cognitive Sciences, CSAIL; Cambridge, MA/USA

Abstract

A visual system cannot process everything with full fidelity, nor, in a given moment, perform all possible visual tasks. Rather, it must lose some information, and prioritize some tasks over others. The human visual system has developed a number of strategies for dealing with its limited capacity. This paper reviews recent evidence for one strategy: encoding the visual input in terms of a rich set of local image statistics, where the local regions grow — and the representation becomes less precise — with distance from fixation. The explanatory power of this proposed encoding scheme has implications for another proposed strategy for dealing with limited capacity: that of selective attention, which gates visual processing so that the visual system momentarily processes some objects, features, or locations at the expense of others. A lossy peripheral encoding offers an alternative explanation for a number of phenomena used to study selective attention. Based on lessons learned from studying peripheral vision, this paper proposes a different characterization of capacity limits as limits on decision complexity. A general-purpose decision process may deal with such limits by "cutting corners" when the task becomes too complicated.

Human vision is full of puzzles. Observers can grasp the essence of a scene in less than 100 ms, reporting with a fair degree of reliability whether it is a beach or a street, whether it contains any animals, and what materials are present [1, 2]. Yet when probed for details, they are at a loss. Change the scene while masking the motion transients, and the observer may have great difficulty determining what has changed, even when the change is quite visible once it has been spotted ("change-blindness", [3, 4]). Human vision is better than the best computer vision systems ever created, yet it is also easily fooled by visual illusions. People can look at a line drawing of a 3D object, and effortlessly understand its shape, yet have difficulty noticing the impossibility of an Escher never-ending staircase. We have difficulty finding our keys, even when they prove quite visible once found and fixated, and look nothing like other items on our desk.

How does one explain this combination of marvelous successes and quirky failures? It perhaps seems unsurprising that these diverse phenomena at present have no unifying explanation. What do they have in common? Certainly, scene perception, object recognition, and 3-D shape estimation require different mechanisms at some stage of visual processing. Nonetheless, might there exist a coherent explanation in terms of a critical stage of processing that determines performance for a wide variety of tasks, or at least a guiding principle for what tasks are easy and difficult?

Attempts to provide a unifying account have explained the failures in terms of the visual system having limited capacity (see [5] for a review). Our senses gather copious amounts of data, seemingly far more than our minds can fully process at once. At any given instant, we are consciously aware of only a small fraction of the incoming sensory input. We seem to have a limited capacity for awareness, for memory, and for the number of tasks we can simultaneously carry out, leading to poor performance at tasks that stress the capacity limits of the system.

A classic example of the limited capacity logic concerns visual search. Suppose a researcher runs an experiment in which observers must find a target item among a number of other "distractor" items. As in many such experiments, the experimenter picks a target and distractors such that individual items seem easy to distinguish. Nonetheless, the researcher finds that search is inefficient, i.e., that it becomes significantly slower as one adds more distractors. Why is search difficult? One can easily discriminate the target from the distractors when looking directly at them. The poor search performance implies that vision is not the same everywhere, or, as Julian Hochberg put it, "vision is not everywhere dense" [6]. If vision were the same throughout the visual field, search would be easy.

By a popular account, the main reason vision is not the same everywhere has to do with attention, in particular selective attention. In this account, attention is a limited resource, and vision is better where the observer attends than where they do not. The visual system deals with limited capacity by serially shifting attention. Some tasks require selective attention, and as a result are subject to the performance limits inherent in having to wait for this limited resource. In the case of difficult search tasks, for instance, the target-distractor discrimination is presumed to require attention, making search significantly slower with increasing number of display items. On the other hand, preattentive tasks do not require attention; they can be performed quickly and in parallel, leading to easy search. Selective attention is typically described as a mechanism that gates access to further visual processing [7, 8, 9] rather than engaging in processing itself. Once the visual system selects a portion of the visual input, perception happens. Throughout this paper, when I refer to selective attention, I mean a gating mechanism. Traditionally, researchers have taken visual search phenomena as evidence that selective attention operates early in the visual processing pipeline, and that correct binding of basic features into an object requires selective attention [10].

Though this account has had a certain amount of predictive power when it comes to visual search, it has been problematic overall [11, 12, 13, 14]. The need for selective attention to bind basic features seems to conflict with: the relative ease of searching for a cube among differently lit cubes [15, 16, 17]; with easy extraction of the gist of a scene [18, 19, 2, 20, 21, 22, 23] and of ensemble properties of sets [24, 25, 26]; and with what tasks require attention in a dual-task paradigm [27].

My lab has argued instead that a main way in which the visual system deals with limited capacity is through encoding its inputs in a way that favors foveal vision over peripheral. Peripheral vision is, as a rule, worse than foveal vision, and often much worse. Peripheral vision must condense a mass of information into a succinct representation that nonetheless carries the information needed for vision at a glance. Only a finite number of nerve fibers can emerge from the eye, and rather than providing uniformly mediocre vision, the eye trades off sparse sampling in the periphery for sharp, high resolution foveal vision. This economical design continues into the cortex: more cortical resources are devoted to processing central vision at the expense of the periphery.

We have proposed that the visual system deals with limited capacity in part by representing its input in terms of a rich set of local image statistics, where the local regions grow — and the representation becomes less precise — with distance from fixation [28]. Such a summary-statistic representation would render vision locally ambiguous in terms of the phase and location of features. Thus, this scheme trades off computation of sophisticated image features at the expense of spatial localization of those features.

One of the main implications of this theory for vision science has been the need to re-examine understanding of visual attention. Most experiments investigating selective attention have had a peripheral vision confound. A number of phenomena previously attributed to attention may instead arise in large part from peripheral encoding.

This paper begins by reviewing both phenomena in peripheral vision and our model of peripheral encoding. It reviews what we have learned about perception, as well as the implications for theories of attention, particularly selective attention. Our understanding of peripheral vision constrains possible additional mechanisms for dealing with limited capacity. In particular, I propose that the brain may face limits on decision complexity, and deal with those limits by performing a simpler version of any too-complex task, leading to poorer performance at the nominal task.

A lossy encoding in peripheral vision

Peripheral vision is susceptible to clutter, as evidenced by the phenomena of visual crowding. Classic crowding refers to greater difficulty identifying a peripheral target when flanked by neighboring stimuli than when it appears in isolation. Crowded stimuli may appear jumbled and uncertain, lacking crucial aspects of form, almost as if they have a textural or statistical nature [29]. Crowding has often been studied with a target identification task, and with a target object flanked by other objects, but it almost certainly affects perception more generally. Crowding points to significant qualitative differences between foveal and peripheral vision. These differences are far greater than the modest differences between foveal and peripheral acuity, and are likely task-relevant for a wide variety of tasks [30]. The phenomena of crowding have been described in detail in a number of recent review papers [31, 32, 33, 34].

My lab has argued that one must control for or otherwise account for the strengths and limitations of peripheral vision before considering explanations based upon visual attention [30, 14, 35]. Otherwise, one risks fundamental misunderstandings about both perception and attention. Whether the paradigm is visual search, change detection, dual-task, scene perception, or inattentional blindness – all tasks whose results have been interpreted in terms of the mechanisms of attention – the often-cluttered stimuli lie at least in part outside of the fovea, and are potentially subject to crowding.

A number of researchers have suggested that crowding results from "forced texture perception," in which information is pooled over sizeable portions of the visual field [29, 36, 31, 32]. Based on these intuitions, we have developed a candidate model of the peripheral encoding that we hypothesize underlies crowding. In this Texture Tiling Model (TTM), originally described in [28], the visual system computes a rich set of summary image statistics, pooled over regions that overlap and tile the visual field. Because of the association with texture perception, we chose as our set of image statistics those from a state-of-the-art model of texture appearance from [37]: the marginal distribution of luminance; luminance autocorrelation; 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 proposal [28, 38] is not so different from models of the hierarchical encoding for object recognition, in which later stages compute more complex features by measuring co-occurrence of features from the previous layer [39, 40, 41, 42]. Second-order correlations are essentially co-occurrences pooled over a substantially larger area.

This encoding scheme provides an efficient, compressed representation. It captures a great deal of information about the visual input. Nonetheless, the encoding is lossy, meaning one cannot reconstruct the original image exactly. We hypothesize that the information maintained and lost by this encoding provides a significant constraint on peripheral processing and constitutes an important and often task-relevant way in which vision is not the same across the visual field.

The proposed lossy encoding has potential implications for virtually all visual tasks. Simply re-examining possible confounds in selective attention studies requires the ability to apply a single model to recognition of crowded peripheral targets, visual search, scene perception, ensemble perception, and dual-task experiments. In order to make predictions for this wide range of stimuli and tasks, one needs a model applicable to arbitrary images.

In addition, critical to our understanding of this encoding scheme has been the use of texture synthesis methodologies for visualizing the equivalence classes of the model. Using these techniques, one can generate, for a given input and fixation, images with approximately the same summary statistics [28, 37, 43, 14, 38]. These visualizations allow for easy intuitions about the implications of the model. Figure 1BC shows two examples synthesized from the image in Figure 1A. Information that is readily available in these synthesized images corresponds to information preserved by the encoding model.

Understanding a model through its equivalence classes is a relatively rare technique in human and computer vision (see [44, 37, 45] for a few notable exceptions). Visualizing the equivalence classes of TTM allows one to see immediately that many of the puzzles of human vision may arise from a single encoding mechanism [38, 28, 43, 14]. Doing so has suggested new experiments and predicted unexpected phenomena [28, 46].

On the other hand, getting intuitions from a low-to-midlevel model by viewing the model outputs is fairly common. Researchers will filter an image to mimic a modeled contrast sensitivity function (CSF), and judge whether the CSF can predict phenomenology (e.g. [47]); they will apply a center-surround filter and judge whether that can predict lightness illusions (e.g. [48]); they will look at a model's predictions for perceived groups, and judge whether they match known perceptual organization phenomena (e.g. [49]).

Furthermore, visualization of the equivalence classes has facilitated the generation of testable model predictions, allowing us to study the effects of this relatively low-level encoding on a wide range of higher-level tasks. Observers view the synthesized images, and perform essentially the original task, whether that be object recognition, scene perception, or some other task [38, 28, 50, 51, 35, 52, 53, 54]. This allows one to determine how inherently easy or difficult each task is, given the information lost and maintained by the proposed encoding. The next section reviews evidence that the proposed encoding can qualitatively – and in a many cases quantitatively – predict a range of visual perception phenomena.

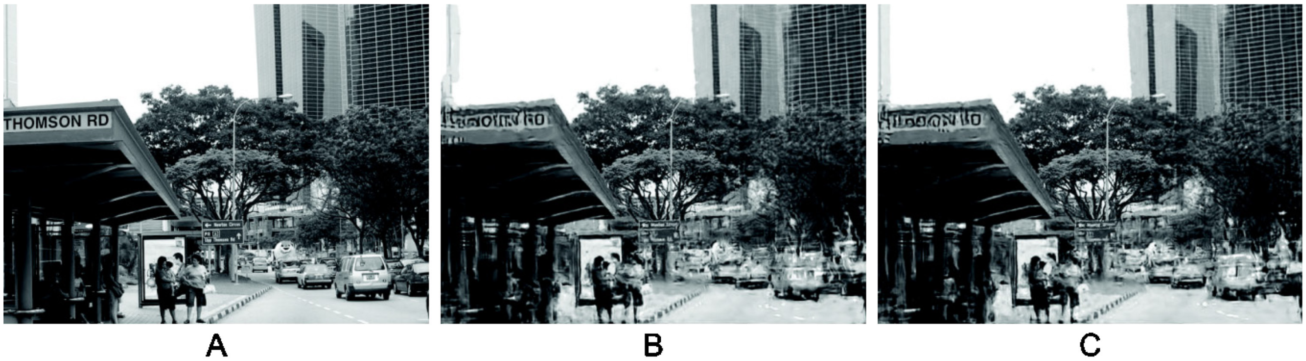

Figure 1. A. Original scene image. B,C. According to the Texture Tiling Model, these images are members of the equivalence class of (A). Details that appear clear in these visualizations are those predicted to be well encoded by the model. TTM preserves the information necessary to easily tell that this is a street scene, possibly a bus stop, with cars on the road, people standing in the foreground, and a building and trees in the background. However, given this encoding, an observer may not be certain of the details, such as the number and types of vehicles, or the number of people.

Limitations of peripheral vision: A factor in many phenomena, and well modeled by TTM

For the last decade, we have worked to re-examine a number of visual phenomena to determine whether peripheral vision was a factor, and whether the encoding modeled by TTM can predict behavioral performance. This includes peripheral object recognition and some of the phenomena associated with the study of visual attention: visual search, scene perception, change-blindness, and dual-task performance.

We have shown that TTM quantitatively predicts performance at a range of peripheral recognition tasks. Balas et al. [28] showed that a local encoding in terms of the hypothesized image statistics can predict identification of a peripheral letter flanked by similar letters, dissimilar letters, bars, curves, and photos of real-world objects. Rosenholtz, et al. [35] and Zhang et al. [52] further demonstrated that this model could predict identification of crowded symbols derived from visual search stimuli. More recently, Keshvari and Rosenholtz [51] have used the same model to explain the results of three sets of crowding experiments, involving letter identification tasks [55], classification of the orientation and position of a crossbar on t-like stimuli [56], and identification of the orientation, color, and spatial frequency of crowded Gabors [57]. In all of these cases, we made predictions based on the information encoded in a single pooling region that included both target and flankers within the critical spacing of crowding. Figure 3A plots some of these results. Note that there are no parameters in the fit of the model to data; model predictions do not merely correlate with behavioral results, but rather quantitatively predict the data.

By incorporating information from multiple, overlapping pooling regions, Freeman and Simoncelli [38] showed that they could predict the critical spacing of crowding for letter triplets. Their pooling region sizes and arrangement were set to make it difficult to distinguishing between two synthesized images with the same local statistics.

Peripheral discriminability of target-present from target-absent patches predicts difficulty of search for a T among Ls, O among Qs, Q among Os, tilted among vertical, and conjunction search [35]. The same is true for search conditions that pose difficulties for selective attention models: cube search vs. search for similar polygonal patterns without a 3-D interpretation [52]. Differences between foveal and peripheral vision are task-relevant for visual search. TTM, in turn, predicts the difficulty of these peripheral discrimination tasks, and thus search (Figure 3B). There is some evidence from search experiments that the model requires additional or different features (e.g. worse encoding of oblique compared to horizontal or vertical lines, and more correlations between different orientations across space). Running TTM has given us some intuitions about how to improve the model.

More recently, with model in hand, we subtly changed classic search displays in ways that should not affect predictions according to traditional selective-attention-for-binding explanations. We changed stroke width, stroke length, or the set of distractors, and correctly predicted whether these changes would make search easier or more difficult [46].

A primary difficulty with early selection accounts has been the ease with which observers can perform many scene tasks. The attentional mechanism that supposedly underlies visual search difficulty has seemed incompatible with the ease with which observers can get the gist of a scene. TTM gives us, perhaps for the first time, a mechanism that can explain both difficult search and easy scene perception. We asked observers to perform a number of navigation-related and other naturalistic scene tasks both at a glance – while fixating the center of the image – and free-viewing. Figure 3C shows predictions of TTM vs. performance at a glance [50]. The prediction is quite good. This graph exaggerates the power of the model of peripheral vision, however, as some tasks are inherently difficult even when free-viewing. Figure 3D compares instead how much more difficult each task is when fixating instead of free-viewing. We see that TTM also does a reasonable job of predicting which tasks are harder when one cannot move one's eyes, i.e. when forced to use extrafoveal vision. Although these are quantitative predictions of scene perception performance, the fit is not parameter-free; in modeling the interaction between multiple pooling regions we chose particular amounts of overlap and density of pooling regions.

Freeman and Simoncelli [38] similarly modeled computations of these image statistics over multiple pooling regions. They adjusted the size of the pooling regions until observers could not tell apart two synthesized images with the same local encoding. They demonstrated that observers have trouble distinguishing between the synthesized "metamers", even when attending to regions with large differences. One can reinterpret this result as showing that they can predict

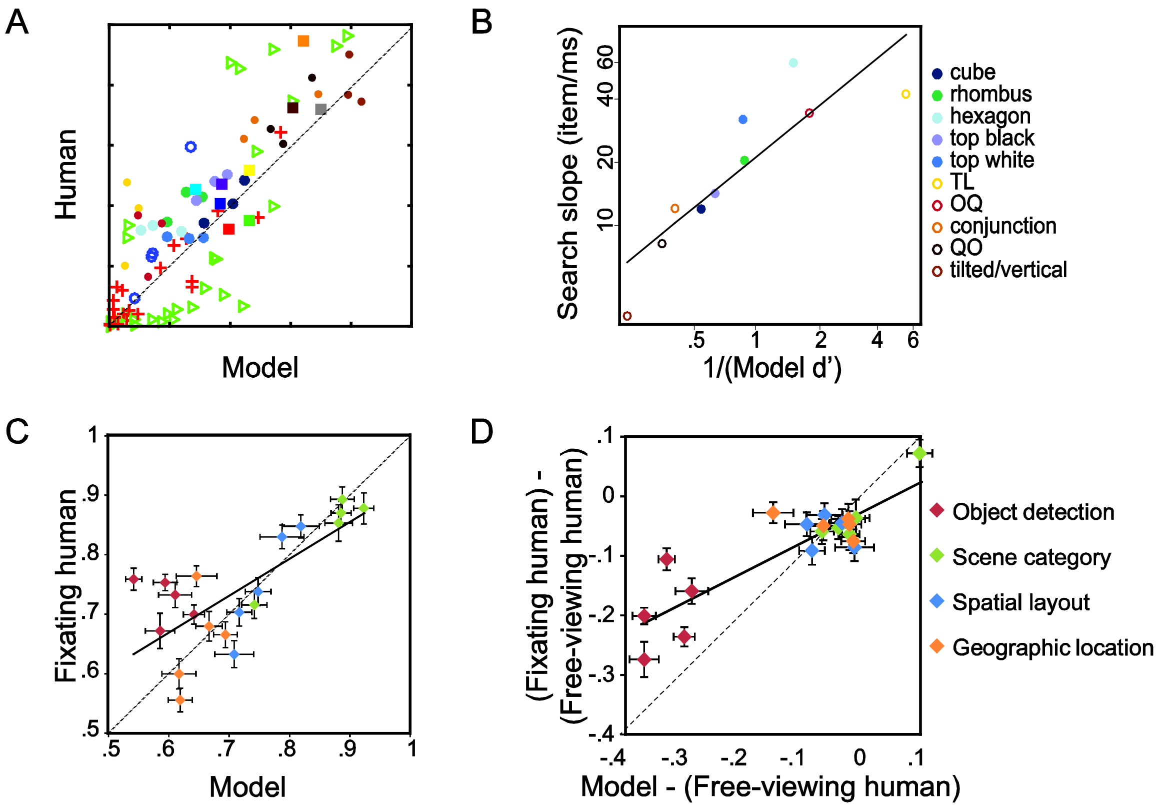

Figure 2. Perfect predictions would lie along the dashed line. Solid line shows best fit between model and data. A. Results from three crowding studies, overlaid (o, triangle, and + from [51], filled circles from [52], squares from [28]). For a wide range of stimuli and tasks, TTM performs well at predicting crowded object recognition. B. TTM predicts search difficulty [52]. C. TTM predicts difficulty at a range of scene gist tasks. D. Many of these scene tasks are only slightly more difficulty when fixating, but object detection is considerably worse. The model correlates well with performance, but underpredicts object detection. [50] For details, please see the original papers.

performance on a different sort of scene task -- telling apart two distorted real-world images.

TTM correlates well with performance at peripheral object recognition, search, and scene perception. However, there is clearly unexplained variance in all of these cases, and room for model improvement. If anything, the model seems to throw away a bit too much information. Human peripheral vision performs somewhat better than predicted, in the case of both crowded object recognition and scene perception (mostly in the case of recognizing a crowded object in the scene).

Easy search tasks do not always correspond to easy dual-tasks (see Figure 2). These results are puzzling if one relies on search and dual-task paradigms to uncover what visual processes require attention, since the two paradigms appear to give different answers [27]. Understanding peripheral vision provides insight into this conundrum [14]. The tasks that are easy in both paradigms are easy in peripheral vision (both behaviorally, and according to TTM). The tasks that are easy in dual-task but hard search tasks simply suffer much more crowding in the search displays. TTM predicts more faithful encoding of a single scene than of an array of scenes, and more faithful encoding of a single colored bar than an array of bars of different orientations and colors. Two of the tasks that are hard in the dual-task paradigm involve distinguishing upright vs. inverted symbols (cube and red-green circle). A single pooling region in TTM would have difficulty distinguishing between upright and inverted, although multiple pooling regions will seldom confuse the two (depending on details such as the size of the stimuli). TTM indicates that these difficult dual-tasks should be more difficult or complex than the easy dual-task conditions. Actually, the upright vs. inverted cube turns out to be a difficult search task when the display is less regular, suggesting the easy search may have resulted from emergent features [14]. For the remaining elements of the matrix (rotated L vs. +, and L vs. T) TTM predicts the search behavior, and we have argued that the difficult dual-task may arise because of

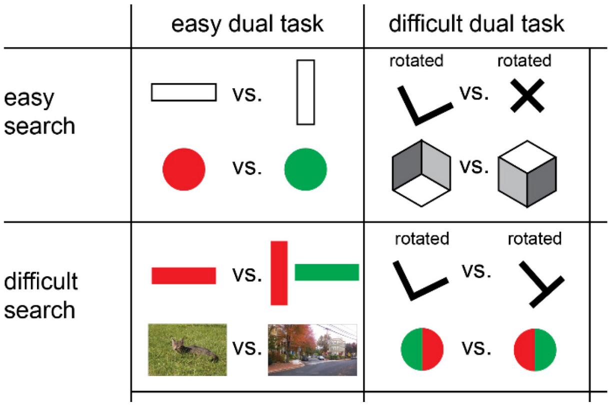

Figure 3. Tasks that seem not to require attention in a visual search paradigm sometimes require attention in a dual-task paradigm. Based on Figure 4A in [27]. TTM helps make sense of these results.

response conflicts with the central T vs. L task. TTM appears to resolve the differences between search and dual-task performance. Note that while TTM provides *insight* into which dual-tasks are easy and hard, it makes clear neither the reason all conditions are easy as a single task, nor the impact of attention, nor precisely why some tasks are easy dual tasks and others are difficult. This paper attempts to address these issues in a later section on decision complexity.

In a phenomenon known as change blindness, observers have difficulty detecting the difference between two similar images, if presented in such a way as to disrupt motion and other transients that would cue the difference [58, 59, 3, 60, 61]. A popular explanation of change blindness suggests that detecting the change requires selective attention to the change in order to perceive and/or encode the changed region in memory (e.g. [3, 62]). Might peripheral vision play a role? We can see hints that a pooling model like TTM predicts change blindness from the visualization of the equivalence classes in Figure 1. While the encoding preserves a great deal of useful information to support getting the gist of the scene, precise details are lost. This may explain why change blindness occurs [38, 14, 63]. Behavioral evidence has also pointed to peripheral vision as a factor; researchers have found eccentricity effects. Observers are worse at detecting changes if they have fixated farther from the changed object, either before or after the change [64, 61, 65]. In pilot work, we asked observers to distinguish between two frames of a standard change blindness stimulus, when both the change and its location were known. Presumably the observers attended to the known location and to the features of the change. We found that the threshold eccentricity at which one can reliably discriminate a known change is predictive of difficulty detecting that change (Sharan et al., under review). Rather than looking for a change merely where one fixates, our results suggest that vision looks for evidence of a change throughout the visual field, possibly in parallel. However, evidence of a change can be weak due to loss of information in peripheral vision, making change detection difficult.

Both behavioral evidence and modeling suggest that peripheral vision plays a greater role than previously thought in a number of phenomena. These phenomena are a subset of those used to study attention, particularly selective attention. The importance of peripheral vision suggests a very different answer to questions like: how can one so easily get the gist of a scene, when one cannot easily search for a T among Ls? Scene tasks may seem to human cognition more complex than judging T vs. L, but due to the nature of encoding in the visual system, the scene tasks are actually simpler, as more information is available. These results by themselves do not preclude the possibility that attentional limits also play a role in the phenomena of search or change blindness. Certainly, attentional limits play some role in dual-task and inattentional blindness phenomena. However, this rethinking certainly calls into question what we have learned about attention, and requires us to reassess what role attention might play.

Revisiting theories of attention

As discussed in the previous section, one way in which the visual system appears to deal with limited capacity is through use of efficient encoding strategies in peripheral vision. This largely static underlying encoding presumably does not vary appreciably with the task. On the other hand, attention, broadly speaking, refers to dynamic, short-term mechanisms by which the brain adapts to the task at hand. Due to the brain's limited capacity, our visual systems cannot automatically perform every possible visual task at the same time. Instead, the brain may concentrate resources on one task in a given moment, and then switch to focus on another task. In attempting to separate the static from the dynamic mechanisms used to deal with limited capacity, what have we learned about the dynamic attentional mechanisms?

Selective attention may not be early

One of the main pieces of evidence in favor of early selection in vision has been visual search [10]. As discussed, the early selection account has been problematic, at best. It struggles to explain easy cube search [15, 16, 17], it struggles to explain the impact on search of small image changes to the target and distractors [46], it seems incompatible with easy scene and set perception [24, 25, 26, 22, 20, 66, 67, 12, 11, 12], and with dual-task results [27]. Physiology early in the visual system (V1/V2), while it shows significant effects of attention, seems not to show the sorts of strong "gating" proposed by early selection accounts [68, 69, 70] or shows it only late (>150 ms) after stimulus onset [70]. Early selection also deviates quite a bit from subjective experience, though reasoning from such experience deserves skepticism. It is unclear whether mere priors on a stable world, memory, and rapid shifts of attention can explain our percept of a rich stable world.

My lab has suggested an alternative explanation for visual search performance that requires no early selection. We have shown that peripheral vision alone can explain the relative difficulty of a range of search tasks [46, 35, 52]. This means, at the very least, that search experiments had a serious confound. Therefore, we cannot neatly interpret search results in terms of the stage at which selective attention operates, thereby calling into question early selection. In addition, our model of encoding in peripheral vision provides a consistent explanation of many puzzling phenomena, including those problematic for an early selection account [14, 35, 52, 50]. In other words, we have not merely called into question the early selection account due to peripheral vision confounds, but have also demonstrated that there exists a viable alternative explanation with more predictive power.

Attention may not in general selectively gate processing

The visual system clearly adapts to perform the task at hand. The question is to what degree that adaptation takes the form of selecting and processing only a portion of the input; is attention merely a gating mechanism? What evidence remains that attention operates by serially selecting a portion of the display for processing?

Many lab tasks explicitly ask observers to selectively process a target. Observers may, for instance, be asked to identify or report a change to only a particular display item among distractors (e.g. [71, 72]), or to track only a subset of the moving items in a display (e.g. [73]). Generally speaking, the visual system is incredibly successful, and we would expect it to do its best to mimic selective attention when doing so is the task. One must question to what degree we can generalize from such tasks to non-selective tasks like search or getting the gist of a scene.

If anything, top-down selective attention tasks provide an example of vision's surprising failures rather than its impressive successes. Lavie et al. [72], for instance, find significant distractor compatibility effects when the task requires observers to respond as to the identity of the target while also remembering a single digit. Perhaps even when the nominal task requires information only about the target it remains a poor evolutionary strategy to ignore all other stimuli [74]. Nonetheless, one can argue that it is odd, if attention generally acts by selecting a portion of the display for further visual processing, that observers are poor at intentionally attending in this way.

In the past, researchers have interpreted both difficult search and difficulty detecting a change as evidence for serial processing, and particularly as evidence for a selective gating mechanism. We have suggested instead that search difficulty arises from losses in peripheral vision. These losses occur even when fully attending. This reopens the door to a largely parallel mechanism punctuated by occasional shifts of fixation to gather more information [75, 76, 35, 77, 78]. These results do not disprove a selective attention story; in general, it is difficult to distinguish between truly parallel processing and very rapid shifts of a serial mechanism, based on behavioral evidence alone. However, it does mean that difficult visual search cannot be used as evidence for selective attention [79, 80, 81, 82, 14, 35, 77].

Similarly, easy scene perception has long been taken as evidence of a more parallel, less selective attention story. Work demonstrating that peripheral vision is a predictive factor in change detection [64, 61, 83, 84] similarly points to more of a parallel mechanism than previously thought. Rather than needing selective attention to a change in order to perceive it, our visual systems appear to utilize parallel processing to look for evidence of a change. However, we are poor at detecting changes at least in part because of loss of information in the periphery. Again, attention in some form may well play a role in change blindness, but we cannot simply interpret change blindness as evidence for selective attention per se.

More recently, my lab has found evidence against a selective gating explanation for the "invisible gorilla" (inattentional blindness) phenomenon. In the original experiment [85], observers are better at noticing the gorilla when counting passes of the team dressed in black than when counting passes of the team in white. Supposedly when selectively attending to the black team, the observer is likely to also select the gorilla, which is colored similarly to black team members. The gorilla thus gains access to higher level processing, allowing the observer to notice it. When selectively attending to the white team, the dissimilar gorilla is unlikely to be selected. We had all observers do a task with the black team, presumably selecting black team players and filtering out white team players. If selective attention to the black team were all that was required to notice the gorilla, noticing rates should be high. Instead, whether or not the observers noticed the gorilla was a function of whether their fixation patterns matched those of people counting black team passes or white team passes. The dependence on fixation suggests a greater role for the strengths and limits of peripheral vision. The lack of dependence on the task-relevant features and regions of the display – where presumably the observer was attending – suggests at best a limited role for selective attention in the invisible gorilla phenomena.

It is also worth noting the mismatch between physiological phenomena and traditional selective attention theory. Physiologically, selective attention effects should look like some sort of suppression of responses to unattended stimuli. To get such effects, however, researchers have had to place multiple stimuli within a single receptive field [70, 8]. Classic selective attention theory instead would predict suppression effects when multiple stimuli appeared anywhere in the visual field.

In summary, recent work has called into question the notion that a primary means of adapting to the task consists of selectively gating some portions of the display for later processing. Some of the evidence in favor of selective attention may have resulted from peripheral vision confounds, and other evidence seems inconsistent with the traditional story.

Should we classify tasks as preattentive vs. requiring attention?

Traditionally, coupled to the notion of selective attention is the idea that some tasks require attentional resources, while others do not. Recent work, however, has called this dichotomy into question.

Researchers have long used visual search tasks to discriminate between which tasks require attention and which are "preattentive". Inefficient search meant that discriminating between the target and distractors required attention, whereas efficient search meant the discrimination could be done preattentively. However, our work calls into question this interpretation, just as it called into question early selection [14, 35, 52]. Search may actually probe peripheral vision, not what does or does not require attention.

The visual search paradigm only implicitly manipulates attention, by stressing resource limits with a task that requires processing multiple items. Dual-task experiments explicitly manipulate attention by asking observers to perform either a single task or two simultaneous tasks. Nonetheless, dual-task experiments may not indicate what tasks do and do not require attention. We have argued that peripheral vision preserves more task-relevant information for performing easy dual tasks than difficult dual tasks [14]. Neither search nor dual-task experiments clearly tell us what judgments require attention.

More generally, researchers have identified few tasks that consistently appear not to require attention. Noticing an oddball item (e.g. a moving item among stationary) or getting the gist of a scene may not require attention [10, 86, 87, 88]. Even these results have been called into question. Detecting a change has long been considered easy if observers have access to a sufficiently salient motion transient. However, Matsukura et al. [89] showed that when performing a secondary task, observers miss changes even when the motion transient is present. Similarly, Cohen et al. [90] have shown that getting the gist of a scene becomes difficult in a dual-task paradigm, so long as the secondary task is sufficiently hard. (See also [91, 92, 93].) It seems that no tasks categorically require no attentional resources. If we abandon the dichotomy of tasks either requiring or not requiring limited attentional resources, where does that leave us?

Constraints on further capacity limits

The study of attention asks, essentially, about the limits on what tasks we can perform at a given moment:

- What the nature of these limits?
- Which tasks run up against a capacity limit and which do not?
- What is the nature of the mechanisms that adapt to the task, and what impact do they have on task performance?

A popular answer has been that of selective attention theory [10, 94]. According to this theory, there are limits on access to higher-level processing; tasks requiring higher-level processing run up against these limits, while tasks requiring only lower-level processing do not; and selective attention deals with these limits by gating access to later stages of processing. However, as discussed above, this account has been problematic. The previous section argued for abandoning notions of early selection, of attention operating primarily by gating access to later processing, and of only a subset of tasks requiring attention.

Alternatively, we have argued that only a limited amount of information can make it through a bottleneck in visual processing; that the visual system deals with that limit by compressing its inputs; and that it does so for all stimuli and tasks – in other words, that this encoding is largely fixed and automatic. However, clearly this is not the whole story. Abundant evidence demonstrates that the visual system encounters additional capacity limits, and deals with these limits by adapting to the task. Behaviorally, dual-task experiments are often harder than single-task (e.g. [27]), an effect not addressed by TTM. Inattentional blindness phenomena demonstrate that performance is better when the observer knows the task [95]. Changing the task (e.g. identify object A instead of B) significantly modifies brain responses [96, 70, 97, 68, 69, 98]. Any viable account of visual processing must explain these effects of task.

Building on this, let us revisit the nature of limited capacity, and of what dynamic mechanisms might adapt to the task. Previous work imposes a number of constraints on the answers to the three questions above. Although it makes sense to describe a theory by answering the questions in the order listed, it is easier to consider the constraints in reverse order: what does previous work say about possible effects of later capacity limits; about which tasks could plausibly face additional capacity limits; and about what might be the nature of those limits?

Constraints on mechanisms that adapt to task

Previous work constrains the nature of additional losses of information due to limited capacity. Losses in peripheral vision already predict the relative difficulty of a range of tasks, including peripheral object recognition [28, 38, 51, 35, 52], visual search [46, 35, 52], and scene perception tasks [50, 14, 38]. Predicted difficulty performing a peripheral task, according to TTM, correlates with difficulty performing that task under dual-task conditions. Difficulty identifying a known, fully attended peripheral change can coarsely discriminate between easy, medium, and hard change detection examples (Sharan, et al., under review). *To be viable, any theory of capacity limits must not (to a first approximation), predict a change in the relative difficulty of those tasks, or we lose predictive power.*

Suppose, for example, that we hypothesize that the visual system encodes unattended portions of the stimulus using a fundamentally different set of summary statistics; attended regions get the rich encoding described by TTM, whereas unattended regions just get the power spectrum, for instance. Any such major change in the encoding would differentially affect some stimuli and tasks compared to others. This would almost certainly change our predictions of the relative difficulty of those tasks, causing us to lose the predictive power we gained by understanding peripheral vision. Mechanisms that cope with additional capacity limits must not make major changes to the visual encoding.

With early selection called into question, and the visual encoding not varying much with task, perhaps we should revisit the possibility that additional capacity limits are late; perhaps even as late as the decision stage. An obvious objection would be that physiology finds effects of task early in visual processing (e.g. [96, 70, 97, 68, 69]. However, this does not immediately preclude late capacity limits, since it remains controversial how to interpret those results. Perhaps some of the early effects of task result from the visual system feeding back a decision -- a hypothesis about the world -- for comparison with the visual input (see [99] for another reinterpretation of attention physiology in terms of effect rather than cause). In other words, physiological effects may follow from successfully performing a particular task, rather than being the mechanism for doing so. Alternatively, mechanisms for dealing with a relatively late capacity limit may nonetheless utilize early processing to some degree.

Which tasks encounter additional capacity limits?

It might seem odd to introduce a dichotomy of tasks that do and do not encounter capacity limits, having just abandoned the dichotomy of whether or not tasks require attention. However, the questions differ, due to a reframing of what we mean by a task. The old question asks whether scene perception, say, never encounters resource limitations, regardless of what other tasks the observer might simultaneously attempt to perform. The new framework treats "scene perception" and "scene perception while tracking multiple objects" as different tasks; the latter may come up against capacity limits while the former does not. This clarifies our interpretation of dual-task experiments: Difficult dual-task performance provides evidence that making the *pair* of judgments encounters capacity limits. It does not provide evidence that the individual component judgments encounter capacity limits.

What insight does previous work provide into what tasks conceivably encounter additional capacity limits? The question is this: for what tasks might performance be *worse* than predicted by the loss of information in peripheral vision?

For many tasks, we do not know the answer. Although losses in peripheral vision are clearly a factor in search, scene perception, and change blindness, this alone does not preclude the possibility that additional capacity limits also play a role. Search tasks may run up against additional limited capacity, or at least hard search tasks may. TTM has predicted relative task difficulty, but not absolute; there is room for, say, a mechanism that makes all difficult search tasks proportionally more difficult. (Additional losses might conceivably also make easy search tasks more difficult than predicted by peripheral vision losses. However, observers so efficiently perform easy search tasks that additional losses must have little effect on those tasks.) The same is true of change blindness. Scene tasks may also run up against additional capacity limits. We did make quantitative predictions of scene task difficulty. However, there were free parameters in the model: the dimensions, density, and arrangement of the pooling regions. Conceivably, TTM threw away too much information, and some of that information may instead be lost due to later capacity limits. Again, any effect on easy scene tasks must be small. In all of these cases, additional modeling could resolve the issue of whether the tasks encounter additional capacity limits.

We do know that harder dual tasks run up against limits in a way that their component single tasks do not. Easier dual tasks do not run up against the same limits. Furthermore, inattentional blindness experiments show that many tasks are easier when the observer knows the task, suggesting that somehow not knowing the task can cause one to encounter capacity limits.

What is limited?

Are we limited in the number of items we can process at once? In the number of tasks we can simultaneously perform? In access to working memory and conscious perception [100]? We should at least try to find a generalizable answer. Proposing a limit in terms of the number of items, for instance, requires a number of qualifiers and clarifications: Do single item tasks never encounter limited resources? Why is search sometimes easy? How many items must one process to recognize a scene? If the limit is on the number of tasks, then what counts as a task, and why are dual tasks sometimes easy? The ultimate answer should not only address the nature of the limits, but also clarify what tasks encounter them, and by what mechanism the visual system deals with them.

Re-examining what tasks might encounter additional capacity limits provides some insight. Here, one can note a general trend that the tasks most likely to encounter additional limits are the hard tasks: hard search, difficult scene tasks, and hard dual tasks. Perhaps the limit is on some resource that makes difficult tasks tractable. For instance, there might exist limited resources for improving the signal-to-noise ratio. Hard discriminations hit limits, and the visual system might adapt by reducing the internal noise limiting one subtask at the expense of others. Alternatively, the visual system might increase the signal by adjusting feature detectors to more precisely measure some features at the expense of others. Unfortunately, this solution does not appear to explain dual-task phenomena. Many dual-task experiments (including those in [27]) adjust the presentation time in order to equalize the difficulty of all single tasks. If all single tasks are equally hard, and capacity limits merely depend on task difficulty, then dual-task conditions should be equally difficult, except perhaps in special cases when the two tasks share features and/or resources.

Why, if one equalizes single task difficulty, are those tasks not equally difficult under dual-task conditions? Equally difficult tasks can be difficult in different ways. The next section discusses the proposal that our visual systems are limited in the *complexity* of the decisions they can make. This hypothesis makes sense of a number of phenomena, and it will soon be testable, by leveraging recent advances in statistical learning.

Proposal: A general limit on decision complexity

Suppose the task was to distinguish between two similar breeds of cats. In some feature space, this hard task might look like one of the examples in Figure 4AB. The feature space, in reality, is certainly high dimensional, but appears here as two dimensions for simplicity. The two classifications might differ behaviorally because the one in Figure 4B might require more training to reach the same level of performance. Once trained, however, the two classification tasks might look quite similar; in the face of observation noise, an ideal observer would perform similarly well in the two cases.

The task in Figure 4B, however, depends on the availability of resources in a way that the task in Figure 4A does not. It can accomplish equally good performance only if it has access to mechanisms capable of implementing a more complicated classification boundary. If one had to perform the task with a single linear classifier, performance would be better in the simple case shown in Figure 4A. Alternatively, the task in Figure 4B may be possible with a linear classifier if one can represent it in a higher-dimensional space; in which case it requires the resources to do so.

Perhaps the visual system has limits on the *complexity* of decision rules. Executive functions may sometimes be unable to construct a sufficiently complex classifier, and as a result, the observer will make errors. The precise nature of the limit might take various forms. For the purposes of discussion, we can think about limits on the number of hyperplanes available to construct the classification boundary. The limit might take other forms, depending upon the nature of the decision-making mechanisms. (If, for instance, the brain implemented classification tasks using center-surround mechanisms operating in some feature space [101], then the limit could be on the number or density of those mechanisms instead.) A limit on decision complexity might exist for good reason, as it would avoid overfitting to sometimes-limited data.

Complexity depends integrally on the underlying encoding. As an initial hypothesis, I suggest thinking of the limit as applying to the feature space at the highest level visual processing. First, arguably that is nearest the decision stage, as appropriate for a limit on decision complexity. Second, the visual system appears to performs a series of hierarchical processing steps precisely to make many real-world tasks simple at higher levels [102]. Finally, even

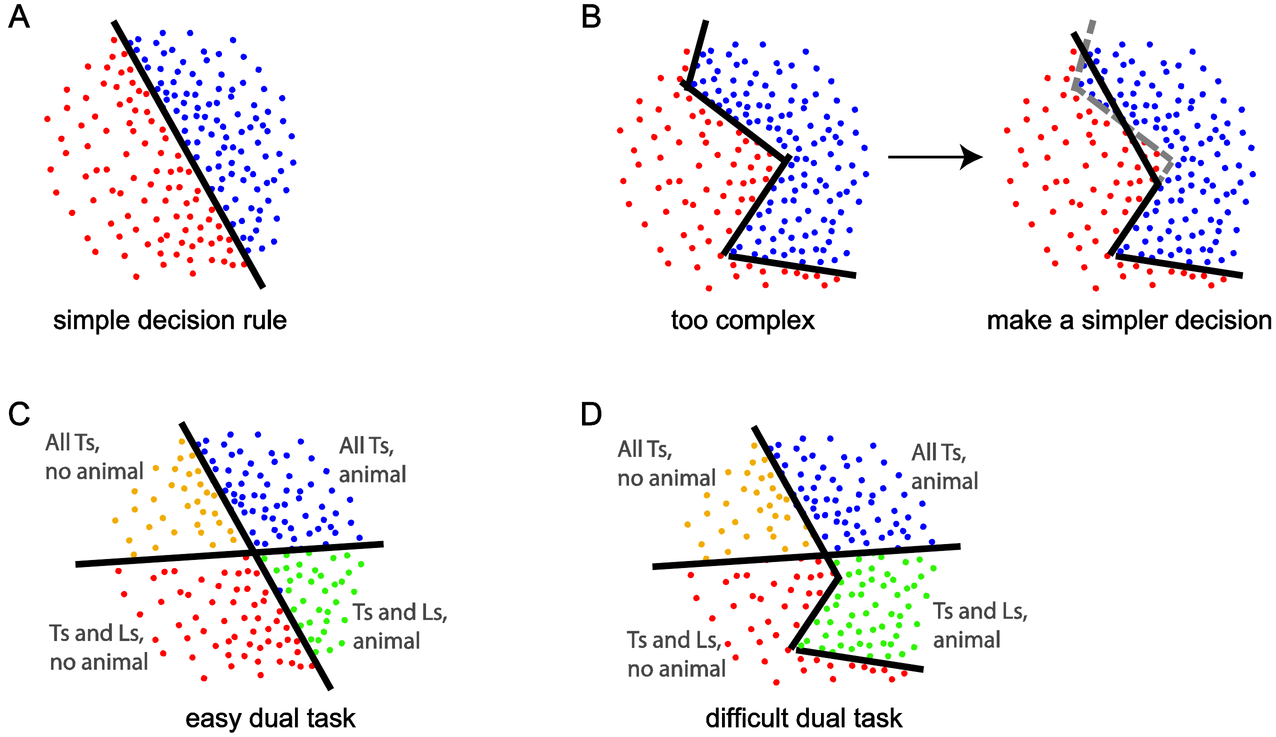

Figure 4. Given the underlying encoding, tasks can be difficult because of similarity between the stimuli to be discriminated (A), or because they require a complex decision rule (B). If the task requires too complex a decision rule, the observer may have to perform a simpler version, "cutting corners" and making errors. Dual-tasks are automatically more complex than single tasks. However, some dual tasks may nonetheless be relatively simple (C), whereas others may be too complex (D), given limited resources.

tasks that seem to require "low-level" information, such as the orientation of a bar, are not obviously simpler given a lower-level encoding. If one does not know the precise location of the bar, early visual encoding provides a soup of responses of feature detectors with different phase and orientation; finding the orientation may well be simpler given a later representation that effectively interprets the bar as an object. Hong et al. [103] find that an inferior temporal population encodes category-orthogonal features such as position and size more explicitly than either a V4 population or a V1 Gabor model.

Tasks that required overly complicated decision rules would encounter complexity limits. To deal with these limits, the brain would have to perform a simpler version of the task. In the example in Figure 4B (left), for instance, a classifier made out of 4 hyperplanes would perform well. However, if the brain only has access to 3 hyperplanes, it would need to carry out a simplified classification. It would literally and figuratively "cut corners" (Figure 4B, right), leading to worse performance at the nominal task. Given enough time, one could perform the original, complex task, by performing a series of simpler tasks; e.g. in a dual-task condition one could first perform one task, and then perform the other.

The brain might make use of a broad range of flexible strategies for simplifying an overly complex decision. A small subset of possible strategies resemble previously proposed attentional mechanisms. Take, as an example, the task of performing two moderately complex tasks at two different locations in the visual field. This joint task might surpass the complexity limit. The brain might simplify by performing one of the two tasks well, at the expense of performance at the other. This sounds something like selective attention to a region of space, except that the suggested mechanism performs the classification, rather than merely gating access to some later stage. One would also expect object-based attentional effects. The representation at the highest level of the visual processing hierarchy arguably developed to make object-based decisions simple [102]. Performing a simpler task might sometimes have the signatures of object-based attentional phenomena. Comprehending the full set of available strategies is difficult without a better understanding of the underlying feature space used for decision-making.

One can immediately see how a limit on decision complexity agrees with a number of the dual-task phenomena. Dual tasks are inherently more complex than their component single tasks. With only a single task, one can recruit all resources (all of the available hyperplanes, for example) to perform the task. For typical component tasks from dual-task studies, observers perform well. However, as shown in Figure 4CD, an observer who has to distinguish between two alternatives for each single task needs, in the dual-task condition, to distinguish between four alternatives. Doing so inherently requires more hyperplanes than either single task alone. However, dual tasks need not all be difficult. If both single tasks are simple enough, the pair may not encounter complexity limits (Figure 4C). Even if the experimenter normalizes the difficulty of all single-task conditions, complicated single tasks will require a complex decision rule, leading to a more difficult dual-task condition. Furthermore, for any given single task, choice of the secondary task can make the dual task easy or difficult.

In fact, TTM has already suggested that difficult dual tasks from [27] are particularly complex. Take for example the two difficult dual-task conditions that require distinguishing between an upright and inverted stimulus (cube, or bisected circle). A single pooling region computing rotationally symmetric statistics cannot distinguish between upright and inverted. However, multiple pooling regions can. In the case of the cube task, one pooling region might identify the top/bottom of the cube, and another could detect a cube without revealing its orientation; together, these pooling regions reveal the orientation of the cube. If the cube were always in the same location, one might be able to do the task using only the outputs of the pooling region that identifies the top of the cube. However, the experiments randomized the cube location on every trial; the location uncertainty makes the task complex.

This explanation perhaps sounds like a tenet of early selection theories: that (to paraphrase) identifying a configuration of features requires limited resources. It is worth pointing out the differences. The features here are likely at a higher level of visual processing, and certainly higher dimensional than in the binding of individual horizontal and vertical features in order to recognize a T [10, 12]. In TTM, as many as 1000 features contribute to perceiving the top/bottom of the cube. Another 1000 features yield the percept of a cube but do not disambiguate its orientation. (Whether or not one needs all 2000 features is of course another question.) Though the proposed capacity limits and mechanisms are quite different, it is not surprising that one can see connections between the old explanations and the new hypothesis; researchers have spent decades gaining intuitions about what tasks are difficult, and we would expect those intuitions to have some truth to them.

One would like, of course, to make testable predictions, for example of dual-task performance. Unfortunately, one cannot easily measure complexity of the decision rule needed for a particular task. A baseline single-task behavioral experiment may not be adequate to predict whether a task will be difficult in a dual-task condition. It will tell us how difficult the single task is, but not how complex it is. We would need to determine complexity in another way. For some novel tasks, it might be possible to examine dependence of performance on the number of training examples, as this dependence will increase with increasing complexity. The upcoming section on testing decision complexity suggests an alternative based on statistical learning techniques.

Decision complexity seems to make sense of other phenomena as well. Human visual encoding likely developed to make scene tasks simple. This may explain why scene tasks are easy in many experimental paradigms, and only become difficult if the experimenter degrades the stimulus in some way, or if the scene task is paired with a particularly complex secondary task [90]. Hard search might or might not be complex because of the need to perform moderately complex tasks at multiple locations, whereas easy search might remain sufficiently simple. If the relative ordering of search tasks according to TTM also serves to order those tasks in terms of their complexity, then limits on decision complexity would maintain the relative ordering, while making hard search tasks harder than predicted by TTM. Whether this is the case, however, remains to be determined. Finally, a limit on decision complexity may explain why, in inattentional blindness, observers perform better when they know the task. If an observer knows the task, they can spend hyperplanes on it. If they do not know the task, why spend excess hyperplanes on an unlikely and unexpected task? Perhaps observers instead use excess hyperplanes to get the gist of the display or the world around them.

Many details remain unspecified: Should one think of complexity limits in terms of hyperplanes, or some other limited resource? What strategies may the visual system employ to simplify a complex decision rule? What is the neural implementation of those strategies? Nonetheless, conceptualizing limited capacity in this different way shows promise at explaining task-based effects.

Do other cognitive systems also encounter limits on decision complexity?

Conceivably, we can take this idea of decision complexity beyond visual perception. The brain might have a similar form of limited capacity in other cognitive systems and for non-visual tasks. For instance, visual memory has limited capacity, and at present lacks a coherent explanation. Perhaps one can make sense of the phenomenology in terms of decision complexity.

Visual short-term memory (VSTM) has appeared distinct from long-term (VLTM) in terms of its capacity; short-term memory appears able to store only 3-4 items [104, 105, 106] (although see, for instance, [107]), whereas long-term memory can store 1000s of objects [108, 109, 110]. However, it is unclear what to make of this capacity difference, as studies of the two types of memory have employed very different stimuli. Short-term memory experiments typically use images containing (e.g.) arrays of simple shapes, and ask whether observers can remember features like the color and orientation of each item (Figure 5B). Long-term memory experiments, on the other hand, utilize photographs of real objects (Figure 5A). Do arrays of colored disks require the same storage capacity as individual toasters and clocks? VSTM also differs from VLTM in the need for an active process to maintain the memory until test. This active storage has identifiable physiological signatures, e.g. sustained activity in frontal and parietal cortices [111, 112]. EEG studies, similarly, have shown sustained activity in the contralateral hemisphere when an observer holds items in short-term memory [113]. Nonetheless, in spite of evidence that the two represent distinct systems, might they suffer from a similar kind of capacity limit?

To assess decision complexity, we first need clarity on what is the memory "task". Researchers often conceptualize memory similarly to storage in a computer: some mechanism stores the information in memory, and some other mechanism is responsible for retrieving that information. Limited storage capacity, as well as storage and retrieval errors, lead to memory errors. From a more decision-theoretic point of view, we can instead think of many visual memory tasks as classification tasks. The observer must distinguish between "seen" and "unseen" stimuli. The experimental subject's job is to derive a classification boundary discriminating between seen stimuli and their best guess as to likely foils used at testing time. (Perhaps also relevant, the subject in short-term memory experiments must do this repeatedly, for many similar displays, whereas in a long-term memory experiment they might construct a single classification boundary and use it for multiple decisions.) This conceptualization of memory tasks differs greatly from the traditional storage/retrieval view, and illuminates the importance of the underlying representation.

To test the viability of memory capacity as resulting from a unified decision-complexity limit, we need a hypothesis for the underlying representation, i.e. the space in which the seen/unseen classification occurs. As a first attempt, suppose that the representation underlying visual recognition also supports and constrains visual memory, both short-term and long-term. This hypothesis makes engineering sense; why develop a domain-general visual representation for recognition, but then use a different representation for memory?

If there exist limits on decision complexity, then one would expect that observers do well on any memory tasks for which a simple decision rule distinguishes between seen (Figure 5, red points) and unseen (blue points), and perform more poorly when the discrimination requires a too-complex rule. Is this a plausible explanation for results in the memory literature? Consider the difference between capacity limits for an array of simple shapes compared to for a series of individual real-world objects. One might think that "simple" stimuli like colored disks would lead to a simpler task than "complex" stimuli like real-world objects. This presupposition should remind the reader of the earlier discussion of whether a T vs. L task is really simpler than a scene classification task. Successive stages of the ventral visual stream are thought to "untangle" the visual representation to enable simple decision rules for real-world tasks such as telling a dog from a tractor [102]. This untangling may also facilitate memory for real-world objects (Figure 5A). However, it potentially comes at the cost of "entangling" less ecologically relevant tasks, such as distinguishing between arrays of different colored disks (Figure 5B). Put another way, to human vision, arrays of colored disks represent a small subset of likely stimuli. They all fall in the category "arrays of colored disks", and the visual system likely encodes them in a way that makes the distance between them very small. (Arguably the

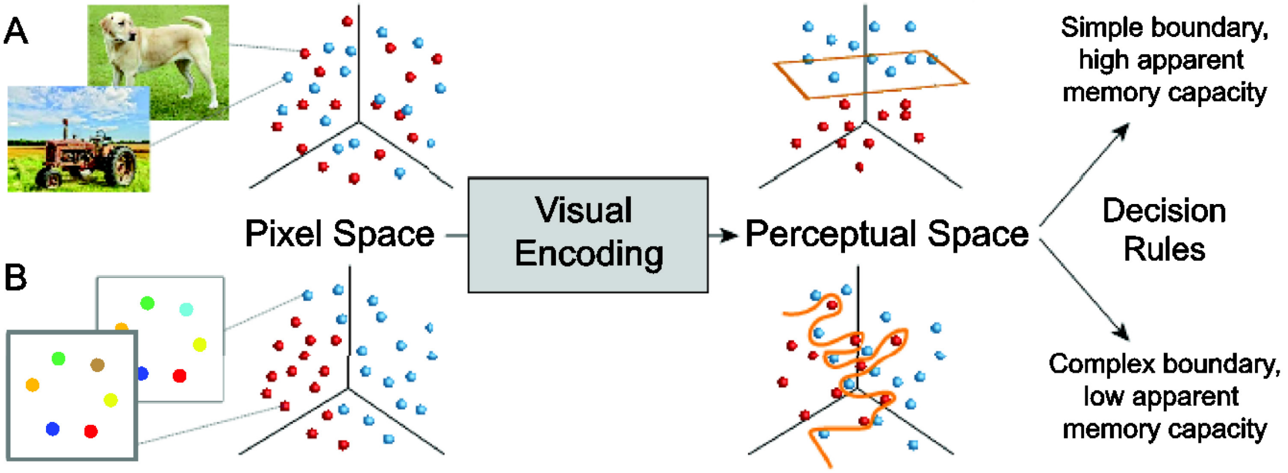
A
B
Pixel Space
Visual Encoding
Perceptual Space
Decision Rules
Simple boundary, high apparent memory capacity
Complex boundary, low apparent memory capacity

Figure 5. A. Discriminating between seen stimuli (red points) and unseen (blue points) may be complex in pixel space, but require a simple decision boundary at a later stage of visual encoding. So long as the boundary remains simple, an observer with a complexity limit can perform the seen/unseen discrimination, leading experimenters to conclude that memory capacity is high. B. Discriminating between similar arrays of disks, on the other hand, may possibly be simple in pixel space, but is likely complex at a later stage of visual encoding. The complexity likely grows with number of disks. A limit on decision complexity would lead to errors for larger number of disks, leading to measurement of a lower memory capacity.

same is true of artificial arrays of real-world objects.) One might expect complexity of the decision to increase as one added more items to the array of colored disks. Discriminating between a particular 2-disk array and all other 2-disk arrays with one changed disk is surely a simpler decision than discriminating between a particular 7-disk array and all other 7-disk arrays with one change. On the other hand, in the case of VLTM for real-world objects, the encoding of toasters and clocks likely makes them far more discriminable.

How one might test decision-complexity theory?

Limits on decision-complexity show initial promise, at least qualitatively, at making sense of both effects of task on perceptual performance, and of visual memory phenomena. To make real quantitative predictions, we need a measure of complexity, and as discussed above it is difficult to measure complexity using behavioral experiments alone. Since decision complexity depends critically on the underlying feature space, going further to test the viability of this theory requires a plausible candidate representation. Given such a representation, the field of statistical learning has developed a number of standard techniques that one can use to evaluate the complexity of a given classification task.

The last few years have seen significant breakthroughs in modeling of visual object recognition. Hierarchical, neurally-inspired models known as convolutional neural nets (CNNs) have suddenly crossed a tipping point. For the first time ever, these "deep" CNNs have allowed computer vision to approach human performance on tasks such as visual object recognition [41], object segmentation [114], and scene recognition [115]. Even more exciting, these CNNs, once well trained on sufficiently difficult visual tasks, with sufficiently variable recognition categories, prove useful for many visual tasks other than the ones for which they were trained [114, 103, 116]. Furthermore, the top layer of one high-performing CNN is highly predictive of neural spiking responses to complex naturalistic images in inferior temporal (IT) cortex, its intermediate layers predictive of responses in V4, and its lowest layers predictive of V1 responses [42].

These results make high-performing CNNs good initial candidates for a general-purpose model of visual representation – essentially, a stand-in for human perceptual processing. For many visual tasks we would need to "foveate" them, to make the representation coarser in peripheral vision and account for crowding [117]. Though CNNs are imperfect, and can respond in ways that do not mimic human performance [118, 119], they may suffice to test decision-complexity theory. Alternatively, one could use fMRI or other physiological data as this candidate visual representation, presuming it were of sufficient resolution and covered enough of the population of neurons.

Next, one needs a measure of decision complexity. Statistical learning theory provides a number of candidates. In statistical learning, classifier complexity is important because of a basic tradeoff: a complex classifier can fit a set of training points well, but one expects that it will make errors on a new set of points, because it is so wiggly – one worries about over-fitting. With a simple classifier, we have more confidence that it will generalize, but it may do a poorer job of fitting the training data. There exist a number of established measures of classifier complexity. One that may be appropriate is the Vapnik-Chervonenkis (VC) dimension [120]. VC dimension specifies the complexity of essentially a class of parametrized classifiers. For a given training set, there is an optimal VC dimension – at low dimensions, the performance may be poor, and at higher dimensions performance slowly declines as a result of over-fitting. Less complex decision rules are those for which the optimal VC dimension is low. Another candidate is sensitivity to classifier regularization. Regularization during classifier training imposes a complexity penalty on classifiers that make use of too many dimensions at once. The sensitivity of the classifier to this penalty provides a measure of the complexity of the classification task. Tasks with low sensitivity are simpler. Alternatively, the number of training examples required to achieve a given level of performance can provide a measure of complexity.

For each task, one needs to define the training and testing examples. Surely one would expect the classifier to be robust to small shifts in the stimulus and small amounts of noise in the representation; this broadens the set of possible training and testing examples beyond the experimenter-defined stimuli. In addition, in the memory task the test set depends upon assumptions about the possible "foils" – the "unseen" stimuli. The experimental subjects may have in mind a different set of foils than the experimenter, leading to different decision complexity and patterns of errors.

Given a candidate feature space and complexity measure, one would like to measure decision complexity for a number of tasks and compare the results. One can represent all training and test examples in the candidate feature space, and for each task determine the complexity of the necessary classifier. (Given a multi-layer architecture like a CNN, one can in fact test the representation at a number of different levels of the processing pipeline.) One can then ask, for instance, whether hard dual tasks lead to higher complexity than easy dual tasks. Similarly, are VLTM experiments with hundreds of items similar in complexity to VSTM experiments with more than 4 items? It should also be possible to try out different strategies for reducing decision complexity, and see whether one can predict observed patterns of errors. In this way, it seems feasible to make testable predictions about the consequences of a capacity limit on decision complexity.

Conclusions

The visual system faces a number of capacity limits. One strategy for dealing with these limits appears to be the efficient encoding of the visual input in a way that becomes less faithful with increasing eccentricity. A model of this encoding, TTM, can explain performance at peripheral object recognition, visual search, and scene perception, and it provides insight into dual-task performance. Furthermore, peripheral task difficulty, measured behaviorally, correlates with search performance and difficulty detecting a change in the absence of a motion transient.

There certainly exist further losses in vision. As processing continues, visual mechanisms continue to pool over ever larger regions of the visual input [121, 122], perhaps leading to further loss of information in the process of gaining invariance to transformations (although see [123]). One might ask, then, why a relatively early encoding has done so well at explaining a range of phenomena. Possibly this arises as an accident of the particular tasks studied. Alternatively, later losses may be more task-specific (e.g. losing lightness information in service of computing lighting-invariant 3-D shape, see [52] for discussion), with an earlier efficient encoding scheme leading to broader and more generalizable effects.

We have learned a great deal from the success of TTM, particularly as regards understanding of visual attention. This paper has argued that selective attention may not operate early in the visual processing pipeline. For that matter, attention may not primarily operate through a selective gating of access to later processing. Perhaps, as well, it is time to abandon the dichotomy of tasks that do and do not require attentional resources. Instead, perhaps we are always attending, i.e. always dynamically adjusting processing in order to perform the current task. In the absence of a nominal task, the visual system may perform some default task, such as getting the gist of a scene, and based on the results of that task, posit and test new hypotheses about the visual world.

In spite of questioning theories of attention, it remains clear that visual processing faces additional capacity limits. Rather than thinking of those as limits on access to higher level processing, perhaps cognitive processing faces general-purpose limits on the complexity of the decision rule. In this formulation, the question is not what tasks do and do not require some resource called attention. Rather, what are the complexity limits, when must one cut corners to reduce complexity, and what are the mechanisms for doing so? These limits might apply not only to vision, but also to other processes such as visual memory. This is not necessarily to say that perception and memory would share the same limited resource, but rather that the capacity limit might take a similar form. Recent successes developing convolutional neural networks to perform complex object and scene recognition tasks may facilitate making testable predictions from this theory.

Acknowledgements
Thanks to Dan Yamins for many fruitful discussions about CNNs, decision limits, memory, and measures of complexity. Portions of the text regarding decision complexity and parts of Figure 5 derive from our co-written grant on understanding visual memory; used with permission. Thanks also the Benjamin Wolfe and Shaiyan Keshvari for useful discussions and helpful comments on the manuscript.

References

[1] M. C. Potter, "Short-term conceptual memory for pictures," *J. Experimental Psychology: Human Learning and Memory,* vol. 2, pp. 502-522, 1976.

[2] S. J. Thorpe, D. Fize and C. Marlot, "Speed of processing in the human visual system," *Nature,* vol. 381, pp. 520-522, 1996.

[3] R. A. Rensink, J. K. O'Regan and J. J. Clark, "To see or not to see: The need for attention to perceive changes in scenes," *Psychological Science,* vol. 8, pp. 368-373, 1997.

[4] D. Simons and D. Levin, "Change blindness," *Trends in Cognitive Sciences,* vol. 1, pp. 261-267, 1997.

[5] J. Driver, "A selective review of selective attention research from the past century," *British Journal of Psychology,* vol. 29, pp. 53-78, 2001.

[6] J. Hochberg, "In the mind's eye," in *Contemporary Theory and Research in Visual Perception*, New York, Holt, Rinehart, and Winston, 1968, pp. 309-331.

[7] A. Treisman and S. Gormican, "Feature analysis in early vision: Evidence from search asymmetries," *Psychological Review,* vol. 95, no. 1, pp. 15-48, 1988.

[8] J. Moran and R. Desimone, "Selective attention gates visual processing in the extrastriate cortex," *Science,* vol. 229, no. 4715, pp. 782-784, 1985.

[9] M. Behrmann, R. S. Zemel and M. C. Mozer, "Object-based attention and occlusion: Evidence from participants and a computational model," *J. Exp. Psych: Human Perception & Performance,* vol. 24, no. 4, pp. 1011-1036, 1998.

[10] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.,* vol. 12, pp. 97-136, 1980.

[11] R. A. Rensink, "Change blindness: Implications for the nature of attention," in *Vision and Attention*, M. R. Jenkin and L. R. Harris, Eds., New York, Springer, 2001, pp. 169-188.

[12] A. Treisman, "How the deployment of attention determine what we see," *Visual Cognition,* vol. 14, pp. 411-443, 2006.

[13] J. M. Wolfe, "Guided Search 4.0: Current progress with a model of visual search," in *Integrated Models of Cognitive Systems*, W. Gray, Ed., New York, Oxford, 2007, pp. 99-119.

[14] R. Rosenholtz, J. Huang and K. A. Ehinger, "Rethinking the role of top-down attention in vision: effects attributable to a lossy representation in peripheral vision," *Frontiers in Psychology,* vol. 3:13, 2012.

[15] J. T. Enns and R. A. Rensink, "Influence of scene-based properties on visual search," *Science,* vol. 247, no. 4943, pp. 721-723, 1990.

[16] J. T. Enns and R. A. Rensink, "Sensitivity to three-dimensional orientation in visual search," *Psychological Science,* vol. 1, no. 5, pp. 323-326, 1990.

[17] J. Sun and P. Perona, "Early computation of shape and reflectance in the visual system," *Nature,* vol. 379, no. 6561, pp. 165-168, 1996.

[18] G. A. Rousselet, J. S. Husk, P. J. Bennett and A. B. Sekuler, "Spatial scaling factors explain eccentricity effects on face ERPs," *Journal of Vision,* vol. 5, no. 10, p. 1, 2005.

[19] M. R. Greene and A. Oliva, "Recognition of natural scenes from global properties: Seeing the forest without representing the trees," *Cogn. Psychol.,* vol. 58, pp. 137-176, 2009.

[20] H. Kirchner and S. J. Thorpe, "Ultra-rapid object detection without saccadic eye movements: Visual processing speed revisited," *Vision Research,* vol. 46, pp. 1762-1776, 2006.

[21] R. VanRullen and S. J. Thorpe, "Is it a bird? Is it a plane? Ultra-rapid visual categorization of natural and artifactual objects," *Perception,* vol. 30, pp. 655-668, 2001.

[22] S. M. Crouzet, H. Kirchner and S. J. Thorpe, "Fast saccades toward faces: Face detection in just 100 ms," *Journal of Vision,* vol. 10, pp. 16.1-16.17, 2010.

[23] M. J. Mace, O. R. Joubert, J. Nespoulous and M. Fabre-Thorpe, "The time-course of visual categorizations: You spot the animal faster than the bird," *PLOS ONE,* vol. 4, no. 6, p. e5927, 2009.

[24] D. Ariely, "Seeing sets: Representation by statistical properties," *Psychological Science,* vol. 12, no. 2, pp. 157-162, 2001.

[25] S.-C. Chong and A. Treisman, "Attentional spread in the statistical processing of visual displays," *Percept. & Psychophys.,* vol. 66, pp. 1282-1294, 2004.

[26] S.-C. Chong and A. Treisman, "Representation of statistical properties," *Vision Research,* vol. 43, pp. 393-404, 2003.

[27] R. VanRullen, L. Reddy and C. Koch, "Visual search and dual tasks reveal two distinct attentional resources," *J. Cogn. Neurosci.,* vol. 16, pp. 4-14, 2004.

[28] B. J. Balas, L. Nakano and R. Rosenholtz, "A summary-statistic representation in peripheral vision explains visual crowding," *Journal of Vision,* vol. 9, no. 12, p. 13, 2009.

[29] J. Y. Lettvin, "On seeing sidelong," *The Sciences,* vol. 16, no. 4, pp. 10-20, 1976.

[30] R. Rosenholtz, "Capabilities and limitations of peripheral vision," *Annual Rev. of Vision Sci.,* vol. 2, no. 1, pp. 437-457, 2016.

[31] D. M. Levi, "Crowding--An essential bottleneck for object recognition: A mini review," *Vision Research,* vol. 48, pp. 635-654, 2008.

[32] D. G. Pelli and K. A. Tillman, "The uncrowded window of object recognition," *Nature Neuroscience,* vol. 11, pp. 1129-1135, 2008.

[33] D. Whitney and D. M. Levi, "Visual crowding: A fundamental limit on conscious perception and object recognition," *Trends in Cognitive Sciences,* vol. 15, no. 4, pp. 160-168, 2011.

[34] H. Strasburger, I. Rentschler and M. Ju"ttner, "Peripheral vision and pattern recognition: A review," *Journal of Vision,* vol. 11, no. 5, p. 13, 2011.

[35] R. Rosenholtz, J. Huang, A. Raj, B. J. Balas and L. Ilie, "A summary statistic representation in peripheral vision explains visual search," *Journal of Vision,* vol. 12(4):14, pp. 1-17, 2012.

[36] L. Parkes, J. Lund, A. Angelucci, J. A. Solomon and J. Morgan, "Compulsory averaging of crowded orientation signals in human vision," *Nature Neuroscience,* vol. 4, pp. 739-744, 2001.

[37] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int. J. Comput. Vis.,* vol. 40, no. 1, pp. 49-71, 2000.

[38] J. Freeman and E. P. Simoncelli, "Metamers of the ventral stream," *Nature Neuroscience,* vol. 14, no. 9, pp. 1195-1201, 2011.

[39] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics,* vol. 36, pp. 193-202, 1980.

[40] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience,* vol. 2, no. 11, pp. 1019-1025, 1999.

[41] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst. 2012*, Lake Tahoe, NV, 2012.

[42] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *PNAS,* vol. 111, pp. 8619-8624, 2014.

[43] R. Rosenholtz, "What your visual system sees where you are not looking," in *Proc. SPIE 7865, Hum. Vis. Electron. Imaging, XVI*, San Francisco, 2011.

[44] D. J. Heeger and J. R. Bergen, "Pyramid-based texture analysis/synthesis," in *Proc. 22nd Annual Conf. on Comp. Graphics & Interactive Techn. (SIGGRAPH '95)*, New York, 1995.

[45] S. C. Zhu and D. Mumford, "A stochastic grammar of images," *Foundations and Trends in Comp. Graphics & Vision,* vol. 2, no. 4, pp. 259-362, 2006.

[46] H. Chang and R. Rosenholtz, "Search performance is better predicted by tileability than by the presence of a unique basic feature," *Journal of Vision,* vol. 16, no. 10, p. 13, 2016.

[47] B. L. Lundh, G. Derefeldt, S. Nyberg and G. Lennerstrand, "Picture simulation of contrast sensitivity in organic and functional amblyopia," *Acta Ophthalmol.,* vol. 59, no. 5, pp. 774-783, 1981.

[48] E. H. Adelson, "Lightness perception and lightness illusions," in *The New Cognitive Neurosciences*, 2nd ed., M. Gazzaniga, Ed., Cambridge, MA: MIT Press, 2000, pp. 339-351.

[49] J. H. Elder, A. Krupnik and L. A. Johnston, "Contour grouping with prior models," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 25, no. 6, pp. 661-674, 2003.

[50] K. A. Ehinger and R. Rosenholtz, "A general account of peripheral encoding also predicts scene perception performance," *Journal of Vision,* vol. 16, no. 2, p. 13, 2016.

[51] S. Keshvari and R. Rosenholtz, "Pooling of continuous feature provides a unifying account of crowding," *Journal of Vision,* vol. 16, no. 3, p. 39, 2016.

[52] X. Zhang, J. Huang, S. Yigit-Elliot and R. Rosenholtz, "Cube search, revisited," *Journal of Vision,* vol. 15, no. 3, p. 9, 2015.

[53] M. N. Agaoglu and S. T. L. Chung, "Can (should) theories of crowding be unified?," *Journal of Vision,* vol. 16, no. 15, pp. 10, 1-22, 2016.

[54] R. G. Alexander, J. Schmidt and G. J. Zelinsky, "Are summary statistics enough? Evidence for the importance of shape in guiding visual search," *Visual Cognition,* vol. 22, no. 3-4, pp. 595-609, 2014.

[55] J. Freeman, R. Chakravarthi and D. G. Pelli, "Substitution and pooling in crowding," *Atten. Percept. Psychophys.,* vol. 74, no. 2, pp. 379-396, 2012.

[56] J. A. Greenwood, P. J. Bex and S. C. Dakin, "Crowding follows the binding of relative position and orientation," *Journal of Vision,* vol. 12, no. 3, p. 18, 2012.

[57] E. Poder and J. Wagemans, "Crowding with conjunctions of simple features," *Journal of Vision,* vol. 7, no. 2, p. 23, 2007.

[58] J. Grimes, "On the failure to detect changes in scenes across saccades," in *Perception (Vancouver Studies in Cognitive Science)*, vol. 2, K. Akins, Ed., New York, Oxford University Press, 1996, pp. 89-110.

[59] G. W. McConkie and C. B. Currie, "Visual stability across saccades while viewing complex pictures.," *J. Exp. Psych.: Human Perception & Performance,* vol. 22, no. 3, pp. 563-581, 1996.

[60] J. K. O'Regan, R. A. Rensink and J. J. Clark, "Change-blindness as a result of 'mudsplashes'," *Nature,* vol. 398, no. 4, p. 34, 1999.

[61] J. K. O'Regan, H. Deubel, J. J. Clark and R. A. Rensink, "Picture changes during blinks: Looking without seeing and seeing without looking," *Visual Cognition,* vol. 7, no. 1-3, pp. 191-211, 2000.

[62] A. Hollingworth and J. M. Henderson, "Accurate visual memory for previously attended objects in natural scenes," *J. of Exp. Psych.: Human Perception & Performance,* vol. 28, no. 1, pp. 113-136, 2002.

[63] A. Oliva and A. Torralba, "Building the gist of a scnee: The role of global image features in recognition," *Prog. Brain Res.,* vol. 155, pp. 23-36, 2006.

[64] J. M. Henderson and A. Hollingworth, "The role of fixation position in detecting scene changes across saccades," *Psychological Science,* vol. 10, pp. 438-443, 1999.

[65] H. L. Pringle, D. E. Irwin, A. F. Kramer and P. Atchley, "The role of attentional breadth in perceptual change detection," *Psychonomic Bulletin & Review,* vol. 8, no. 1, pp. 89-95, 2001.

[66] R. A. Rensink, "Seeing, sensing, and scrutinizing," *Vision Research,* vol. 40, pp. 1469-1487, 2000.

[67] G. A. Rousselet, O. Joubert and M. Fabre-Thorpe, "How long to get to the "gist" of real-world natural scenes?," *Visual Cognition,* vol. 12, no. 6, pp. 852-877, 2005.

[68] C. J. McAdams and J. H. R. Maunsell, "Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4," *J. of Neuroscience,* vol. 19, no. 1, pp. 431-441, 1999.

[69] B. C. Motter, "Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli," *J. Neurophys.,* vol. 70, no. 3, pp. 909-919, 1993.

[70] S. J. Luck, L. Chelazzi, S. A. Hillyard and R. Desimone, "Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex," *J. Neurophys.,* vol. 77, pp. 24-42, 1997.

[71] M. I. Posner, "Orienting of attention," *Q. J. of Exp. Psych.,* vol. 32, pp. 3-25, 1980.

[72] N. Lavie, A. Hirst, J. W. de Fockert and E. Viding, "Load theory of selective attention and cognitive control," *J. Exp. Psych.: General,* vol. 133, no. 3, pp. 339-354, 2004.

[73] Z. W. Pylyshyn and R. W. Storm, "Tracking multiple independent targets: Evidence for a parallel tracking mechanism," *Spatial Vision,* vol. 3, no. 3, pp. 1-19, 1988.

[74] S. S. Shimozaki, M. P. Eckstein and C. K. Abbey, "Comparison of two weighted integration models for the cueing task: Linear and likelihood," *J. of Vision,* vol. 3, no. 3, p. 3, 2003.

[75] W. S. Geisler and K.-L. Chou, "Separation of low-level and high-level factors in complex tasks: Visual search," *Psychological Review,* vol. 102, pp. 356-378, 1995.

[76] W. S. Geisler, J. S. Perry and J. Najemnik, "Visual search: The role of peripheral information measured using gaze-contingent displays," *Journal of Vision,* vol. 6, no. 9, pp. 858-873, 2006.

[77] A. H. Young and J. Hulleman, "Eye movements reveal how task difficulty moulds visual search," *J. Exp. Psych.: Human Perception & Performance,* vol. 39, no. 1, pp. 168-190, 2013.

[78] G. J. Zelinsky, "A theory of eye movements during target acquisition," *Psychological Review,* vol. 115, no. 4, pp. 787-835, 2008.

[79] M. P. Eckstein, "The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing," *Psychological Science,* vol. 9, pp. 111-118, 1998.

[80] J. Palmer, C. T. Ames and D. T. Lindsey, "Measuring the effect of attention on simple visual search," *J. of Exp. Psych.: Human Perception & Performance,* vol. 19, no. 1, pp. 108-130, 1993.

[81] M. Carrasco, D. L. Evert, I. Chang and S. M. Katz, "The eccentricity effect: Target eccentricity affects performance on conjunction searches," *Perception & Psychophysics,* vol. 57, pp. 1241-1261, 1995.

[82] M. Carrasco, T. L. McLean, S. M. Katz and K. S. Frieder, "Feature asymmetries in visual search: Effects of display duration, target eccentricity, orientation, & spatial frequency," *Vision Research,* vol. 38, pp. 347-374, 1998.

[83] G. J. Zelinsky, "Eye movements during change detection: Implications for search constraints, memory limitations, and scanning strategies," *Perception & Psychophysics,* vol. 63, no. 2, pp. 209-225, 2001.

[84] R. E. Parker, "Picture processing during recognition," *J. Exp. Psych.: Human Perception & Performance,* vol. 4, no. 2, pp. 284-293, 1978.

[85] D. J. Simons and C. F. Chabris, "Gorillas in our midst: Sustained inattentional blindness for dynamic events.," *Perception,* vol. 28, pp. 1059-1074, 1999.

[86] F. F. Li, R. VanRullen, C. Koch and P. Perona, "Rapid natural scene categorization in the near absence of attention," *PNAS,* vol. 99, no. 14, pp. 9596-9601, 2002.

[87] S. Otsuka and J. Kawaguchi, "Natural scene categorization with minimal attention: Evidence from negative priming," *Perception & Psychophysics,* vol. 69, no. 7, pp. 1126-1139, 2007.

[88] G. A. Rousselet, M. Fabre-Thorpe and S. J. Thorpe, "Parallel processing in high-level categorization of natural images," *Nature Neuroscience,* vol. 5, no. 7, p. 629, 2002.

[89] M. Matsukura, J. R. Brockmole, W. R. Boot and J. M. Henderson, "Oculomotor capture during real-world scene viewing depends on cognitive load," *Vision Research,* vol. 5, no. 1, pp. 546-552, 2011.

[90] M. A. Cohen, G. A. Alvarez and K. Nakayama, "Natural-scene perception requires attention," *Psych. Science,* vol. 22, no. 9, pp. 1165-1172, 2011.

[91] A. M. Larson, T. E. Freeman, R. V. Ringer and L. C. Loschky, "The spatiotemporal dynamics of scene gist recognition," *J. Exp. Psych: Human Perception & Performance,* vol. 40, no. 2, pp. 471-487, 2014.

[92] A. Mack and J. Clarke, "Gist perception requires attention," *Visual Cognition,* vol. 20, no. 3, pp. 300-327, 2012.

[93] G. A. Rousselet, S. J. Thorpe and M. Fabre-Thorpe, "Processing of one, two, or four natural scenes in humans: the limits of parallelism," *Vision Research,* vol. 44, no. 9, pp. 877-894, 2004.

[94] D. Broadbent, Perception and Communication, London: Pergamon Press, 1958.

[95] A. Mack and I. Rock, Inattentional blindness, Cambridge, MA: MIT Press, 1998.

[96] L. Chelazzi, E. K. Miller, J. Duncan and R. Desimone, "Responses of neurons in macaque area V4 during memory-guided visual search," *Cereb. Cortex,* vol. 11, no. 8, pp. 761-772, 2001.

[97] C. J. McAdams and J. H. R. Maunsell, "Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4," *J. Neurosci.,* vol. 19, no. 1, pp. 431-441, 1999.

[98] L. Reddy, N. G. Kanwisher and R. VanRullen, "Attention and biased competition in multi-voxel object representations," *PNAS,* vol. 106, no. 50, pp. 21447-21452, 2009.

[99] R. J. Krauzlis, A. Bollimunta, F. Arcizet and L. Wang, "Attention as an effect not a cause," *Trends in Cognitive Sciences,* vol. 18, no. 9, pp. 457-464, 2014.

[100] P. E. Dux and R. Marois, "The attentional blink: A review of data and theory," *Attention, Perception, & Psychophysics,* vol. 71, pp. 1683-1700, 2009.

[101] S. L. Franconeri, G. A. Alvarez and P. Cavanagh, "Flexible cognitive resources: Competitive content maps for attention and memory," *Trends in Cognitive Sciences ,* vol. 17, no. 3, pp. 134-141, 2013.

[102] J. J. DiCarlo and D. D. Cox, "Untangling invariant object recognition," *Trends in Cognitive Sciences,* vol. 11, no. 8, pp. 333-341, 2007.

[103] H. Hong, D. L. K. Yamins, N. J. Majaj and J. J. DiCarlo, "Explicit information for category orthogonal properties increases along the ventral stream," *Nature Neuroscience,* vol. 19, pp. 613-622, 2016.

[104] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychological Review,* vol. 63, pp. 81-97, 1956.

[105] S. J. Luck and E. K. Vogel, "The capacity of visual working memory for features and conjunctions," *Nature,* vol. 390, pp. 279-281, 1997.

[106] N. Cowan, "The magical number 4 in short-term memory: A reconsideration of mental storage capacity," *Behavioral and Brain Sciences,* vol. 24, pp. 887-185, 2001.

[107] W. J. Ma, M. Husain and P. M. Bays, "Changing concepts of working memory," *Nature Neuroscience,* vol. 17, no. 3, pp. 347-356, 2014.

[108] L. Standing, "Learning 10,000 pictures," *Q. J. of Exp. Psych.,* vol. 25, pp. 207-222, 1973.

[109] T. F. Brady, T. Konkle, G. A. Alvarez and A. Oliva, " Visual long-term memory has a massive storage capacity for object details," *PNAS,* vol. 105, no. 38, pp. 14325-14329, 2008.

[110] J. L. Voss, "Long-term associative memory capacity in man," *Psychonomic Bulletin & Review,* vol. 16, pp. 1076-1081, 2009.

[111] K. Sakai, J. B. Rowe and R. E. Passingham, "Active maintenance in prefrontal area 46 creates distractor-resistant memory," *Nature Neuroscience,* vol. 5, pp. 479-487, 2002.

[112] J. J. Todd and R. Marois, "Capacity limit of visual short-term memory in human posterior parietal cortex," *Nature,* vol. 428, pp. 751-754, 2004.

[113] E. K. Vogel and M. G. Machizawa, "Neural activity predicts individual differences in visual working memory capacity," *Nature,* vol. 428, pp. 748-751, 2004.

[114] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *arXiv preprint arXiv:1311.2524 (2013),* 2013.

[115] B. Zhou, J. Xiao, A. Lapedriza, A. Torralba and A. Oliva, "Learning deep features for scene recognition using Places database," *Advances in Neural Information Processing Systems,* vol. 27, 2014.

[116] A. S. Razavian, H. Azizpour, J. Sullivan and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," *arXiv preprint arXiv:1403.6382 (2014),* 2014.

[117] T. Poggio, J. Mutch and L. Isik, "Computational role of eccentricity dependent cortical magnification," *arXiv preprint arXiv:1406.1770 (2014),* 2014.

[118] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199 (2013),* 2013.

[119] C. Zhang, S. Bengio, M. Hardt, B. Recht and O. Vinyals, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530 (2016),* 2016.

[120] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability & Its Applications,* 1971.

[121] E. Kobatake and K. Tanaka, "Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex," *J. Neurophys.,* vol. 71, pp. 856-867, 1994.

[122] B. A. Wandell and J. Winawer, "Computational neuroimaging and population receptive fields," *Trends in Cognitive Sciences,* vol. 19, no. 6, pp. 349-357, 2015.

[123] N. C. Rust and J. J. DiCarlo, "Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT," *J. Neuroscience,* vol. 30, no. 39, pp. 12978-12995, 2010.