Vision Research 181 (2021) 10–20

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Vision Sciences Society Young Investigator Award 2018

The meaning and structure of scenes

Melissa Le-Hoa Võ *

Department of Psychology, Johann Wolfgang-Goethe-Universität, Frankfurt, Germany

ARTICLE INFO

Keywords:
Scene grammar
Attention
Scene perception
Search
Prediction hierarchies
Anchors

ABSTRACT

We live in a rich, three dimensional world with complex arrangements of meaningful objects. For decades, however, theories of visual attention and perception have been based on findings generated from lines and color patches. While these theories have been indispensable for our field, the time has come to move on from this rather impoverished view of the world and (at least try to) get closer to the real thing. After all, our visual environment consists of objects that we not only look at, but constantly interact with. Having incorporated the meaning and structure of scenes, i.e. its "grammar", then allows us to easily understand objects and scenes we have never encountered before. Studying this grammar provides us with the fascinating opportunity to gain new insights into the complex workings of attention, perception, and cognition. In this review, I will discuss how the meaning and the complex, yet predictive structure of real-world scenes influence attention allocation, search, and object identification.

1. Introduction

The bulk of studies in vision science has used highly abstract, artificial tasks and simplified stimuli. While this approach has proven essential in forming our understanding of the basic principles of vision, the field has reached a point at which we need to highlight the need for a new type of ecological perspective (Gibson, 1979), one that builds upon well-controlled laboratory research while additionally seeking to understand how we make sense of and interact with our actual environment (Hayhoe, 2017; Hayhoe & Ballard, 2005; Hayhoe & Rothkopf, 2010). Luckily, our visual world might be more complex in nature than most laboratory stimuli, but it rarely presents itself to us as chaotic. Instead — similar to words as part of sentences — objects as part of a scene tend to be structured in a rule-governed way. This "grammar" seems to be the source of efficient scene understanding, object recognition, and goal-directed behavior (Biederman, Mezzanotte, & Rabinowitz, 1982; Draschkow & Võ, 2017; Võ & Wolfe, 2013a, 2013b; Võ, Boettcher, & Draschkow, 2019).

In the 1980s, Biederman et al. (1982) demonstrated that objects violating our generic knowledge of the world are more difficult to identify when presented briefly in an inconsistent scene context. Although this initial perceptual account of incongruent objects was later challenged (e.g., by controlling for response biases, Henderson & Henderson, 1998), Biederman's initial taxonomy that described various relations between objects and their surroundings still inspires work today. Biederman suggested that "something roughly analogous to what may be needed to account for the comprehension of sentences is required to account for the speed and accuracy of the comprehension of scenes never experienced before." (Biederman, 1976). Accordingly, the terms "*semantics*" and "*syntax*" have been used to describe object-scene relationships that determine *what* objects should be *where* within a scene, respectively (Biederman et al., 1982; Võ & Henderson, 2009; 2011; Võ & Wolfe, 2013a). We have modified this initial conceptualization such that objects that do not fit the overall meaning of a scene (e.g. a piece of cheese in the bathroom) are referred to as *semantic* violations, while we consider objects that are semantically consistent, but structurally unexpected (e.g. toilet paper in the shower) as *syntactic* violations.

2. Objects in context

When objects repeatedly appear together within certain contexts (e.g., a pot on a stove in a kitchen), these experienced regularities will create predictions which are stored as part of our scene grammar. This allows us to not only rapidly process and interpret the current visual input, but also predicts other elements of a visual scene that might not

* Address: Scene Grammar Lab, Department of Cognitive Psychology, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 6, 60323 Frankfurt am Main, Germany.
E-mail address: mlvo@psych.uni-frankfurt.de.
URL: https://www.scenegrammarlab.com/.

https://doi.org/10.1016/j.visres.2020.11.003
Received 23 August 2019; Received in revised form 31 October 2020; Accepted 3 November 2020
Available online 8 January 2021
0042-6989/

yet be fully visible (see Bar, 2004 for a review). Perception is therefore not merely a reactive but a proactive process and object contexts play a crucial role in sharpening our predictions.

Already in the early 1980s, Biederman et al. (1982) were able to show that objects semantically consistent with the scene category they were presented within are identified faster and more precisely than objects that do not fit the scene category. Even when they are only briefly presented, objects placed on task-irrelevant background images are named less accurately if they are inconsistent with these background scenes and vice versa (see also Davenport & Potter, 2004).

Beyond explicit naming of objects, event-related potentials (ERPs) — that are present in the electroencephalogram (EEG) signals — are able to indicate an observer's covert response to such stimulus manipulations. Particularly in the domain of language, EEG recordings have been used to distinguish semantic and syntactic processing, since these two types of processing have shown to elicit two distinct types of ERP responses: Roughly speaking, an N400 signals semantic violations (for a review see Kutas & Federmeier, 2011), while a P600 marks inconsistent syntactic structure.

We have previously extended this finding from language processing to scene perception by measuring ERPs while observers saw semantically or syntactically invalid objects within scenes in order to look for neural markers that distinguish these two types of processing. In fact, we found a clear dissociation between ERP signatures of semantic and syntactic scene processing (see Fig. 1; Võ & Wolfe, 2013a): As predicted by the sentence processing literature as well as pioneering work in the domain of visual perception by Ganis and Kutas (2003), semantic object-scene inconsistencies produced negative deflections in the N300/N400 time window (blue line in Fig. 1) similar to the N400 responses seen in language processing (see also Demiral, Malcolm, & Henderson, 2012; Mudrik, Lamy, & Deouell, 2010; Mudrik, Shalgi, Lamy, & Deouell, 2014), while syntactic inconsistencies elicited a late positivity resembling the P600 found for syntax manipulations (see Fig. 1, red solid lines; see also Cohn, Jackendoff, Holcomb, & Kuperberg, 2014 for semantic and syntactic processing in visual narratives). Recently, EEG activity recorded during observation of actions has shown that similarly action *content* versus action *structure* also elicit differential brain responses reminiscent of dissociations found in the domains of language or music (e.g., Maffongelli et al., 2015; Patel, 2003). Critically, even though context-sensitive processes such as the N400 interact with perceptual stages of object recognition (Draschkow, Heikel, Võ, Fiebach, & Sassenhagen, 2018), this late neural response is largely independent of low level stimulus properties, and is reflected in different event-related potentials (Truman & Mudrik, 2018).

Interestingly, extreme syntax violations — i.e. physical violations such as a floating toaster (see dotted red line in Fig. 1) — failed to produce a P600 effect. A floating toaster might be so at odds with our expectations regarding scenes that no reanalysis is triggered. This is in line with findings using linguistic stimuli where extremely ungrammatical sentences also fail to elicit a P600 response due to a lack of reanalyzing the sentence (see Hopf, Bader, Meng, & Bayer, 2003).

One should not take such similarities between linguistic, action, music, and/or scene processing as evidence for identical grammatical processing across domains. However, it may be that there are shared cognitive mechanisms that govern these percepts.

3. Local and global context effects on object processing

While it seems clear that scene contexts influence the processing of objects, it has remained unclear which "ingredients" of a scene are sufficient to modulate such processing driving the reported consistency effect? One can imagine scene context influencing object processing in at least two different ways: *global* effects stemming from broad properties of a scene or more *local* effects due to the immediate surroundings of a particular object. We have known since Molly Potter's seminal work in the 1970s, that the semantic content or gist of a scene can be accessed from images presented at rates up to 10 per second (Potter, 1975). Diagnostic colors have also been shown to mediate scene recognition (Oliva & Schyns, 2000). Moreover, statistical and structural cues extracted from very brief (e.g. 19 ms, masked) exposures allow for above-chance semantic categorization of scenes (natural/urban: Greene & Oliva, 2009). This can be done without the need to segment and identify the objects embedded in a scene (for a review see Oliva & Torralba, 2007). As a consequence, non-selective information extracted

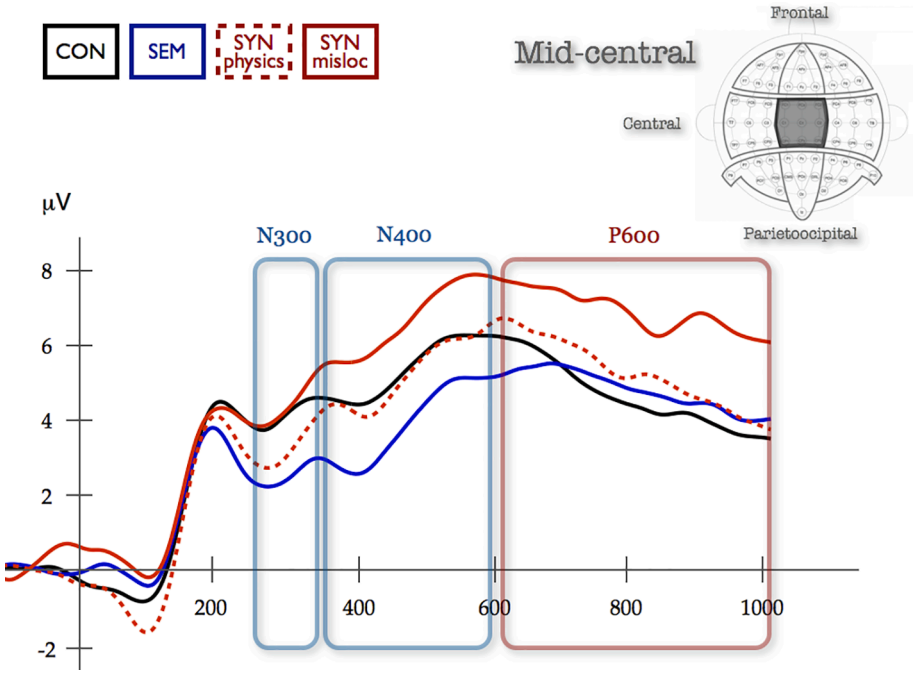

Fig. 1. Grand-average event-related potential (ERP) waveforms recorded from the midcentral region and computed for the consistent control, inconsistent-semantics, inconsistent-syntax/extreme-violation, and inconsistent-syntax/mild-violation conditions. The time windows for the N300, N400, and P600 ERP components are highlighted (taken from Võ & Wolfe, 2013a).

from the initial glimpse of a scene can serve as a quick and powerful source of information that guides search within that scene (Castelhano & Henderson, 2007; Larson & Loschky, 2009; Larson, Freeman, Ringer, & Loschky, 2014; Võ & Henderson, 2010; for a review see Wolfe, Võ, Evans, & Greene, 2011).

We have recently tested the influence of global image statistics by investigating how so-called scene textures as introduced by Portilla and Simoncelli (n.d.) , modulate object processing. These textures have the feature that they retain a collection of global statistical measures – based on basic visual features – similar to a source image, but discard spatial layout information (see Fig. 2 left vs. middle column; Lauer, Cornelissen, Draschkow, Willenbockel, & Võ, 2018). Replicating previous studies, we found that objects presented on semantically consistent scenes were identified significantly better than if they were presented on inconsistent scenes. Objects presented on scene textures had a similar, but much weaker effect. When recording ERPs, a pronounced mid-central negativity in the N300/N400 time windows was triggered by inconsistent relative to consistent objects on scenes. When we presented these objects on scene textures, inconsistent objects resulted in similar brain responses characterized by slightly weaker N300/N400 components. These results imply that the low-level features of scenes at least contribute to the semantic processing of objects in complex real-world environments.

In addition to these global scene properties allowing for rapid extraction of scene gist and activation of scene knowledge, more local information surrounding an object might play a role in its processing as well. For instance, objects themselves have been shown to activate scene context representations (for a review, see Trapp & Bar, 2015), which then again can prime other objects within a scene. However, a more direct route to activate object representations probably includes co-occurring objects, i.e. seeing a toothbrush will allow you to guess that a tube of toothpaste rather than a tube of mustard is likely nearby (see also Mack & Eckstein, 2011). More specifically, through a lifetime of seeing objects in specific configurations we have acquired knowledge regarding the spatial positioning of not only the target, but also distractor objects which could speed object search and facilitate perception. For instance, Gronau and Shachar (2015) presented objects either contextually related (e.g. lamp and desk) or unrelated (keys and apple) and found that such contextual manipulations on the object level increased visual detail encoding into LTM at very short presentation durations. However, these were isolated objects on white background. For other objects within a cluttered scene to influence the processing of a critical object, these other objects would need to be identified first, which – in most cases – would be too time consuming for the local context to exhibit the rapid consistency effects demonstrated in previous studies. That said, the processing of a bathroom sink might help recognizing the tube of toothpaste even within a glimpse, but I will discuss these "special" types of objects in scenes in Section 8.

In sum, both local and global information will most likely contribute to the efficiency of not only object recognition, but also object search in naturalistic scenes. Thus identifying the key ingredients of scene contexts that affect object processing and studying the complex interplay of local and global information processing in scenes will yield interesting new insights into the efficiency of human perception.

4. Attention allocation in real-world scenes

While the rapid interplay of local and global information is fascinating, we usually do not see the world in brief glimpses. Instead, we gather information by moving our eyes to different parts of a scene, accumulating information as we go. One of the key questions in the field of scene perception has been what determines where and when we attend during scene viewing. Computer vision approaches have been influential in their attempts to answer these questions. Computational models of saliency, for instance, initially did a decent job in predicting where people will look using exclusively bottom-up feature contrasts (Itti & Koch, 2000; Koch & Ullman, 1985), especially when objects as mid-level features are taken into account (Malcolm & Shomstein, 2015). Incorporating neurophysiological constraints on saccade programming, MASC – a model of attention inspired by the superior colliculus – can predict where people look during free viewing of scenes or during search (Adeli, Vitu, & Zelinsky, 2016). While most models predicting gaze have focused on predicting *where* people look, both the CRISP (Nuthmann, Smith, Engbert, & Henderson, 2010) and more recently the LATEST model have been developed with the aim to predict *when* the eyes move (Tatler, Brockmole, & Carpenter, 2017).

What all these models have in common is that the information used to model gaze stems from mainly low-level visual features, but there are

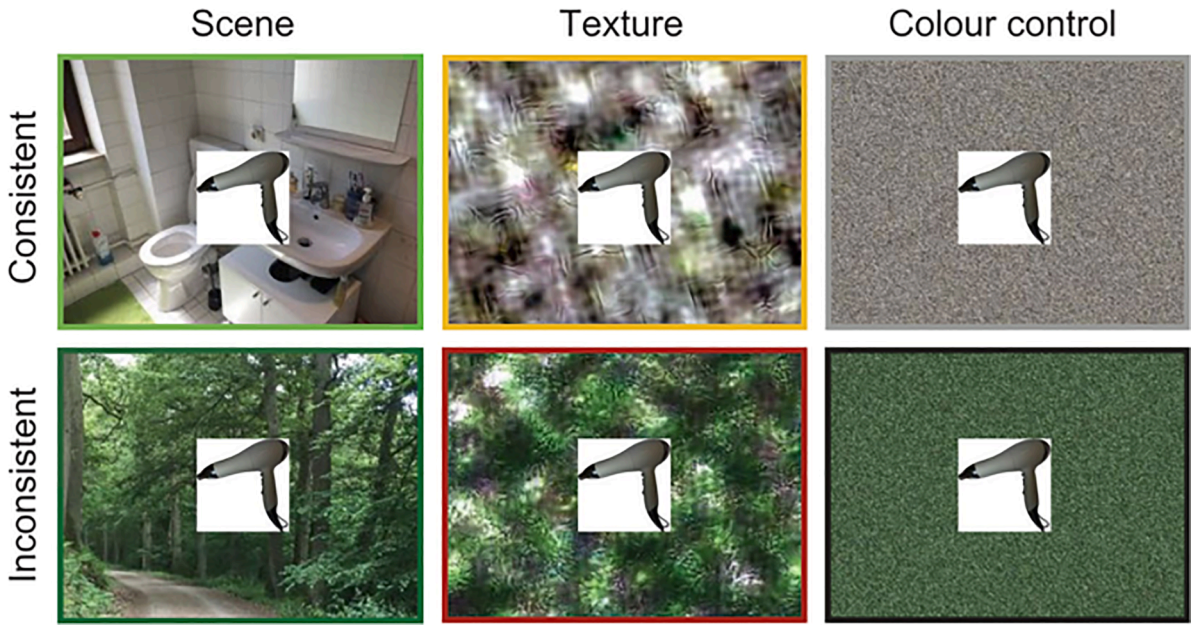

Fig. 2. Example of a stimulus set with objects superimposed on either scenes (left column), scene textures that preserve similar summary statistics as compared to the full forest scene but discards global shape information (middle), or color controls (right) (taken from Lauer et al., 2018).

limits of saliency as an explanatory tool (e.g. Henderson, Brockmole, Castelhano, & Mack, 2007; Henderson, Malcolm, & Schandl, 2009; Tatler, Hayhoe, Land, & Ballard, 2011). When searching for your phone, the visual information of your phone or your surroundings will initially not be an important guidance factor. Instead, you tend to start by searching your episodic memory, revisiting places where you usually leave your phone or revoking memory traces on where you last saw it, leaving little to no room for the influence of an object's bottom-up saliency.

It has been shown that attention and eye movements are less bound to target low-level features and more to meaningful units, like objects (e.g. Nuthmann & Henderson, 2010; Stoll, Thrun, Nuthmann, & Einhäuser, 2015). Latest work using deep neural networks to identify objects, has impressively shown that simply by including features pre-trained on object recognition into a model, its performance on predicting observers' eye movements during scene viewing can be boosted (for an overview of the contribution of low-and high-level contribution to fixation predictions see Kümmerer & Gatys, 2017).

Knowledge regarding co-occurring objects as well as probable scene regions where objects tend to be found provides strong contextual cues during real-world search (e.g., Mack & Eckstein, 2011; Neider & Zelinsky, 2005; Torralba, Oliva, Castelhano, & Henderson, 2006), even when objects are semantically inconsistent with their surroundings implying an independent influence of both local and global information on attention allocation in scenes (Castelhano & Heaven, 2011; Koehler & Eckstein, 2017). Castelhano and Heaven (2010) were able to disentangle the differential contributions of target features, gist and scene details for attentional guidance showing that both scene context and target features speeded search and that previewing visual details of a scene improved search guidance beyond its gist alone.

The superiority of top-down knowledge over bottom-up saliency in guiding visual attention has been demonstrated often and in many compelling ways (e.g. Eckstein, Koehler, Welbourne, & Akbas, 2017; Henderson et al., 2009; Vo & Henderson, 2009; Võ & Henderson, 2010). But the best example is my 3-year-old daughter who knows exactly where to look for the cookies specifically hidden from view in the kitchen, no target features needed.

Henderson and colleagues (Henderson, Hayes, Peacock, & Rehrig, in press) have started directly comparing the influences of meaning and image salience on attentional guidance in real-world scenes (see also Hwang, Wang, & Pomplun, 2011; Wu, Wick, & Pomplun, 2014 for guidance by semantics). Here, the spatial distribution of semantic features in a scene is represented as a meaning map (Henderson & Hayes, 2017), which are generated from crowd-sourced responses of participants who rate the meaningfulness of a large number of scene patches drawn from various scenes. They showed that when the correlation between meaning and saliency was statistically controlled, only meaning uniquely accounted for variance in attention.

Eye movements have frequently been used as an implicit indicator of attention allocation in space signaling semantic processing, for instance when something is amiss. Since the seminal work by Loftus and Mackworth (1977), longer looking times on semantically inconsistent compared to consistent objects have been replicated many times reflecting increased attentional demands (e.g., Bonitz & Gordon, 2008; Cornelissen & Võ, 2016; de, Graef, Christiaens, & Gd'Ydewalle., 1989; Henderson, Weeks, & Hollingworth, 1999; Loftus & Mackworth, 1977; Öhlschläger & Võ, 2016; Underwood, Templeman, Lamming, & Foulsham, 2008; Vo & Henderson, 2009). Also syntactic inconsistencies lead to longer dwell durations (e.g. Henderson et al., 1999; Öhlschläger & Võ, 2016; Vo & Henderson, 2009; see also the SCEGRAM database for a highly controlled set of images containing objects that undergo various types of semantic and syntactic violations provided by Öhlschläger & Võ, 2016). The eyes even get "stuck on semantics" when observers are told to search for Ts amongst Ls overlaid on task-irrelevant background scenes that contain semantically inconsistent objects (Cornelissen & Võ, 2016). This implies that the relationship between an object and its scene is processed automatically. Thus, even task-irrelevant semantic mismatches within a background scene can have an impact on where we look.

While there is agreement on the fact that people look longer at semantically inconsistent objects, it has been a matter of debate as to whether such inconsistencies can also be processed in and thus attract attention to the visual periphery (Becker, Pashler, & Lubin, 2007; Coco, Nuthmann, & Dimigen, 2020; de et al., 1989; LaPointe & Milliken, 2016; Underwood, Humphreys, & Cross, 2009; Vo & Henderson, 2009; Võ & Henderson, 2011). These conflicting results seem to be due to a mixture of stimulus properties (e.g., the saliency and size of the critical objects), task (e.g., visual search vs. change detection), and overall differences in global properties of the scenes used (e.g. cluttered photographs vs. sparser 3D rendered images). Larger scale meta analyses of the stimulus materials used combined with regression-based approaches incorporating stimulus differences across studies could potentially elucidate on this matter. Until then it should be safe to state that both scene semantics and syntax heavily influence attention and ongoing eye movement behaviour during real-world scene viewing.

5. Memory for objects in scenes

We have seen that our memory for objects in scenes, i.e. past experiences that we use to generate predictions, seems to be key when interacting with real-world environments. Particularly when searching for an object, we tend to guide our search by a mixture of memories: Generic knowledge about what objects tend to be where – scene grammar – as well as more specific memories about that particular scene stored in episodic memory. When studying the influence of memory on search, the stimulus material commonly used contains arbitrary items such as oriented bars or rotated Ts amongst upright Ls. Interestingly, the classic contextual cueing studies impressively demonstrated that even such seemingly meaningless "scenes", could be learned to an extent that speeded search without participants being explicitly aware about the fact that they had repeatedly seen the same target-distractor arrangements (Chun & Jiang, 1998). Observed search benefits result from associating the location of a particular target exemplar with a particular search array allowing for more efficient allocation of attention to a subset of the visual display that most likely contains the target item. Thus, episodic memory for previous exposures to a scene — even when implicit — can improve search.

Searching through real-world scenes is a very different thing, though. Here, target-context associations are much more abstract and are governed by semantic expectations (Brockmole & Le-Hoa Vo, 2010). Moreover, there has been a bulk of studies now demonstrating that once we move to more realistic stimuli, we seem to have massive memory for images of both objects and scenes that goes way beyond our capacity of memorizing meaningless items (e.g., Konkle, Brady, Alvarez, & Oliva, 2010; but see Cunningham, Yassa, & Egeth, 2015; Draschkow, Reinecke, Cunningham, & Võ, 2019 for the importance of what type of encoding and which memory test is used). Visual details of previously fixated (and therefore likely attended) objects in naturalistic scenes can be stored in LTM for hours or even up to days (for a review see Hollingworth, 2006). Even incidental fixations on objects during search improve subsequent recognition memory (e.g., Castelhano & Henderson, 2005; Võ, Schneider, & Matthias, 2008).

These studies support the conclusion that looking at an object provides the observer with considerable information regarding its appearance and position within the scene. Thus, it seems reasonable to believe that repeatedly searching through the same visual scene should increase guidance by episodic memory to speed search. However, in a study by Wolfe and colleagues (Wolfe, Alvarez, Rosenholtz, Kuzmova, & Sherman, 2011a), participants repeatedly searched the same scenes for various different objects with surprisingly little reduction in search efficiency across searches. Võ and Wolfe (2012) replicated these findings tracking eye movements and argued that in real-world scenes the ability

to use guidance by scene semantics and syntax diminishes the role of episodic memory because when contextual guidance is strong it is faster to just search again than reactivating possibly faulty episodic memory representations (but see Hollingworth, 2012). If this were true, one would predict that episodic memory should become more useful if other sources of scene guidance are lacking. Võ and Wolfe (2013b) put this to a test by presenting participants with search displays containing inconsistently placed objects (i.e. syntactic guidance was misleading) and found that now episodic memory did indeed boost search.

On the other hand, once an object has been found, searching for it a second time is significantly speeded (Hollingworth, 2012; Võ & Wolfe, 2012, 2013b; Wolfe et al., 2011a). This implies that looking *at* an object versus looking *for* an object has differential effects on the object representations encoded and stored into memory (Võ & Wolfe, 2012). However, actual memory performance for searched targets had not been explicitly tested in any of these studies. Therefore we directly compared memory performance of target objects incidentally encoded during search in scenes with that created by having another group of participants intentionally memorize the same objects (Draschkow, Wolfe, & Vo, 2014). We found that memory recall was actually substantially better for searched objects that were incidentally encoded than for objects that had been intentionally memorized. This "search superiority effect" remained stable even when gaze durations on the critical objects were kept equal across conditions. Note that the mere act of finding an object cannot explain this memory benefit for searched items, since when the same experiment was performed with random object displays instead of depictions of real-world scenes, search superiority disappeared (Josephs, Draschkow, Wolfe, & Võ, 2016). Thus, in addition to aiding object search, scene grammar may also help create a beneficial scaffolding that promotes memory for objects that one has previously looked for.

6. Cognitive development of scene knowledge

It is clear that we have a vast body of scene knowledge that influences our perception. However, we do not enter this world knowing where objects should be, but over time have learned the rules of our world by constantly interacting with it. Some of this knowledge seems to be there from early age. There is a large corpus of work showing that knowledge regarding basic physical laws and number concepts, for instance, is already well developed in infants (for reviews see Hespos & vanMarle, 2011; Spelke, Breinlinger, Macomber, & Jacobson, 1992). Baillargeon (1987) demonstrated that infants as young at 3.5 months of age look significantly longer at events where a rotating screen appears to pass through space occupied by a box as when it behaves physically expected, indicating that infants at that age already possess an understanding of physical laws. Spelke et al. (1992) further showed that infants as young as 2 months looked longer at a ball that did not stop when it came in contact with a solid wall. Thus, "core knowledge" systems seem to be in place at a very early stage. Moreover, eye tracking studies have shown that infants as young as 4 months can predict the reappearance of occluded objects (Johnson, Amso, & Slemmer, 2003).

In contrast, understanding the relationship between an object and its meaningful context seems to occur later in life. Ratner and Myers (1981), for instance, asked children to name objects that would fit certain rooms in a dollhouse. They found that major conceptual changes took place between ages two and three in that 2-year-olds produced fewer core items (objects that we would call "anchor objects") than the 3- and 4-year-olds. Moreover, the consistency effect – i.e. longer gaze durations on semantically inconsistent objects – known from studying adults was also demonstrated in two-year olds, but only when attention was already directed to the critical objects by their high visual saliency (Helo, van Ommen, Pannasch, Danteny-Dordoigne, & Rämä, 2017).

In a recent study (Öhlschläger & Võ, 2020), we investigated the behavioral responses of 72 two- to four-year-olds in two tasks that either measured scene knowledge directly by asking them to furnish a dollhouse or indirectly by observing their eye movements when viewing scene photographs that included semantically inconsistent objects. The consistency effect as we know it from studying adults was evident only in children older than three years. Interestingly, the differences in first-pass dwell durations between consistent and inconsistent objects were due to shorter processing times, i.e. faster disengagement, for consistent objects reflecting stronger predictions for objects in their familiar context/location as children grow older. This reduction of first-pass dwell times to consistent objects correlated with the dollhouse performance measure of scene knowledge. These results imply that scene-related predictions can effectively influence both implicit and explicit behavior at the latest by the age of four years allowing optimized attention allocation in scenes.

While eye movements provide an easily accessible, implicit measure of contextual effects on object processing, measuring ERPs such as the N400 component in response to semantic context violations has been used as a possibly more sensitive indicator for modulations of semantic processing in adults (e.g. Lauer et al., 2018; Lauer, Willenbockel, Maffongelli, & Võ, 2020). In a recent study, we found that already 2-year-old toddlers showed modulations of the N400 component in response to objects presented semantically inconsistent contexts (Maffongelli, Öhlschläger, & Võ, 2020). This implies that by the age of two, toddlers might have already developed scene semantic knowledge to a degree that allows them to detect purely visual, semantic object-scene inconsistencies. While that might not yet become evident in overt eye movements, the brain seems to already pick up on these semantic irregularities.

Early deficits in semantic processing — as seen in the absence of N400 components in response to semantic picture-word violations — have been shown to correlate, for instance, with enhanced risk for the development of specific language impairment (SLI) (Friedrich & Friederici, 2006). Both the development of language and the development of scene knowledge are essential for everyday life and both require the integration of new experiences into existing knowledge structures. It is therefore possible that the acquisition of language and the refinement of scene knowledge go hand in hand, drawing on similar cognitive resources. To this date, only few studies have investigated this relationship during development (e.g., Helo et al., 2017; Öhlschläger & Võ, 2020; Saarnio, 1990). Clearly further work is needed to elucidate whether the development of language and perception really draw on shared knowledge structures.

7. Real-World environments

In order to understand how we represent our world it is important to investigate cognition under more realistic settings in which participants are free to move around the environment and perform more natural tasks (Gibson, 1979; Hayhoe, Shrivastava, Mruczek, & Pelz, 2003; Kingstone, Smilek, & Eastwood, 2008; Malcolm, Groen, & Baker, 2016; Tatler et al., 2011, but see Holleman, Hooge, Kemner, & Hessels, 2020 for a critical review of the term "ecological validity").

In general, studying perception in real-world environments shifts the focus away from the properties of the stimulus toward a consideration of the behavioral goals of the observer (like navigation, obstacle avoidance, or grasping) as well as the behavior's metabolic costs. For instance, Gilchrist, North, and Hood (2001) compared search for arbitrary objects in 2D displays (symbols) with search for arbitrary objects in 3D environments (arrays of film canisters) and found an increased use of memory when body movements were involved. And there is further evidence that acting on objects influences our memory for them (e.g. Draschkow & Võ, 2016; Harman, Humphrey, & Goodale, 1999; Trewartha, Case, & Flanagan, 2015). In Draschkow and Võ (2016), for instance, we showed that interacting with objects during a search task within 3D environments modulated memory for the object's identity and position as a function of whether it was task-relevant versus irrelevant.

However, active behavior does not always seem to enhance memory representations. After active exploration of a virtual environment, recall and recognition memory were neither improved for object identity nor location memory (Brooks, Attree, Rose, Clifford, & Leadbetter, 1999).

More recently, we used virtual reality (VR) to let participants construct their own environments from scratch, either consistent to their scene grammar or in a manner they considered inconsistent (Draschkow & Võ, 2017). We tested observers' memory in two ways: explicitly, by having them rebuild some rooms with all objects and their locations or implicitly, by having them repeatedly search through the formerly built rooms. With this setup, we were able to demonstrate that contextual violations, even those that were created by the observers themselves, can impede both explicit memory and search performance. As an additional indicator of memory involvement, we also looked at the development of search efficiency over the course of repeated searches within one scene and found that compared to 2D searches (Võ & Wolfe, 2012; Wolfe, 1998), searches in this 3D environment do become faster over time (see also Draschkow & Võ, 2016; Helbing, Draschkow & Võ, 2020; see also Li, Aivar, Kit, Tong, & Hayhoe, 2016). Moreover, we replicated the "search superiority effect" in that we found superior memory for searches versus memorized objects also in fully immersive, 3D environments (Helbing et al., 2020).

While we cannot and should not do without well-controlled and simplified laboratory studies, I strongly believe that we additionally need to design experiments in as realistic scenarios as possible. This would allow us to critically question which findings really translate to the real world and which might be artefacts of paradigms too far from what we do in real life. But more importantly, in order to study the meaning and structure of scenes in their full complexity, we need to investigate attention, perception, and cognition in scenarios that at least mimic our daily experiences, while creating new methods and techniques that allow us to better analyse, visualize and make sense of this increasingly rich data.

8. Spatial hierarchies of objects in scenes

What makes a scene and how are scenes organized? Is a depiction of an office room as much of a scene as a close-up of a desk? And how are predictions that we have about *what* objects should be *where* organized? There are different ways to approach these questions. Most of the studies – including our own - that investigated the types of predictions we have regarding a scene have presented observers with images that depict violations of their expectations, e.g. semantic vs. syntactic violations. In Draschkow and Võ (2017), we opted for a new paradigm and utilized virtual reality technology to overcome previous constraints by investigating how people would build their own environments. That is, by providing them with empty rooms that they were instructed to equip according to how they felt it was "grammatically correct", we were able to "watch" scenes being built from scratch according to the scene grammar of our participants, something that would be utterly difficult to do in the real world (unless you work at a moving company and ask your movers to furnish initially empty rooms). Looking at how and in which order the participants handled the different objects, we could literally see how large objects that are generally hard to move were placed first, followed by arranging smaller objects around them. As can be seen in Fig. 3, global objects were moved early in each trial and seemed to define the space, which can then be filled according to a more fine-grained grammar with other local objects in relation to the global ones. We have termed these global objects "anchors" (Boettcher, Draschkow, Dienhart, & Võ, 2018), because they seem to play an important role in "anchoring" the position of other objects within a scene (e.g. "I first need to place the toilet in order to position the toilet paper next to it.").

These data demonstrate experimentally that not all objects are created equal. For instance, it has previously been shown that some objects tend to be more important for scene categorization, i.e. are more "diagnostic", than others (Biederman, 1981; Biederman et al., 1982; Friedman, 1979; Greene, 2013). More importantly, especially in indoor scenes, objects tend to cluster around anchors forming meaningful subgroups or – as we have started to call them - "phrases", e.g. individual sink vs. shower vs. toilet phrases that together make up the larger bathroom (see Fig. 4). Knowledge about this partitioning of a scene obviously will be helpful in making your search more efficient: When looking for the shampoo, you can quickly disregard the sink and the toilet phrases, focusing your search on only the behaviorally relevant sub unit, here the shower phrase. Similarly, this knowledge of a scene's hierarchical composition will speed object recognition: When you see something standing on top of the stove in the corner of your eye, the predictions regarding what types of objects usually rest on top of this anchor will quickly narrow down the list of possible objects and let you identify the blurry object as a pot.

What now do we mean by "anchor objects" and how do they differ from merely being large, diagnostic objects? In addition to often being prominent objects that are diagnostic for a scene, e.g. the shower in the

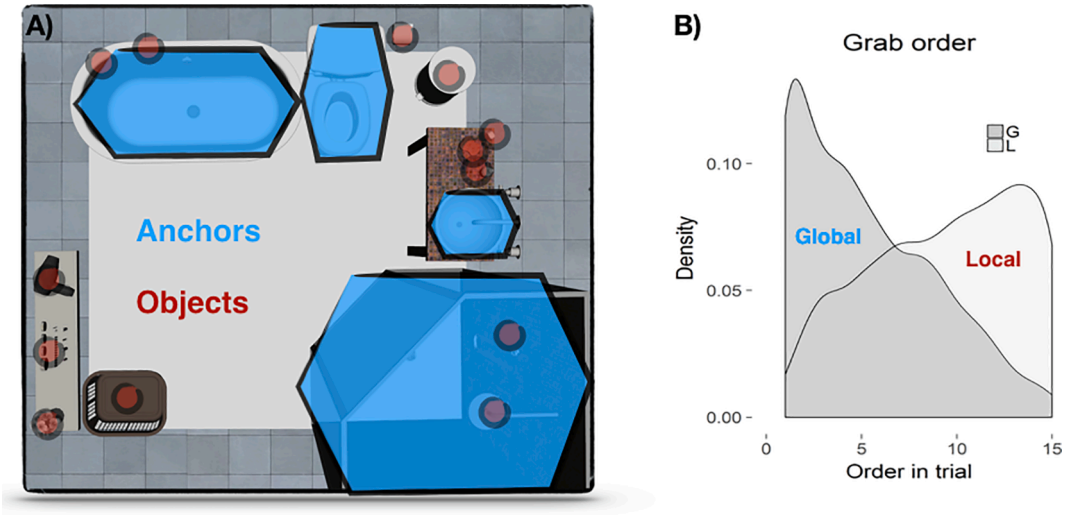

Fig. 3. The left figure shows a Birdseye depiction of a bathroom that was constructed in a manner participants considered consistent. Local objects were placed relative to anchors. The right graph depicts computed density estimates (y-axis) for first object grabs during a trial (x-axis) as a function of Object type (global vs. local) showing that global objects (anchors) were handled early within a trial.

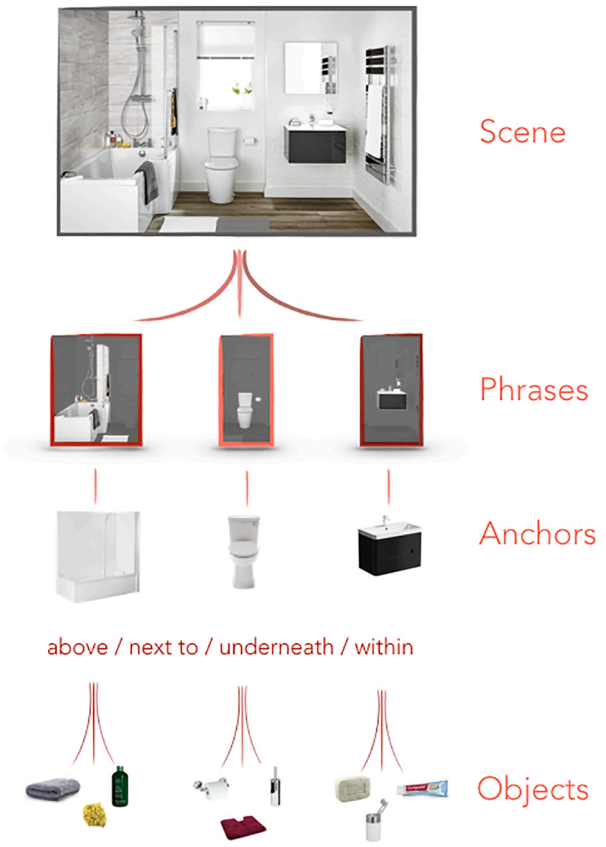

Fig. 4. Proposed hierarchical organization of a bathroom scene that includes three phrases that again consist of one anchor each (e.g. a shower, a toilet and a sink) that predict the locations of other objects (e.g. the shampoo is *in* the shower, the toothbrush *on top of* the sink, the toilet paper *next to* the toilet, etc.).

bathroom or the stove in the kitchen, the most important feature of anchors is that other objects within the same phrase have defined spatial relations relative to their anchors. That is, diagnostic objects mainly are able to tell us *what scene* we are in, while anchors over and above that provide us with predictions regarding *where objects* are. e.g. the shampoo *in* the shower, the pot *on* the stove, the lamp *beside* the bed (see Draschkow & Võ, 2017). A toilet brush, on the other hand, might be diagnostic of the scene category bathroom, but will not exactly tell us where the toilet paper is.

Moving away from a binary – often intuitively based – definition of anchors, we have tried to operationalize their otherwise vague conceptualization by using four concrete determinants (Boettcher et al., 2018; Võ, Boettcher, & Draschkow, 2019): 1) the frequency in which objects appear together, 2) the distance between objects, 3) the variance of the spatial location, and 4) the clustering of objects within scenes. Out of these we formulated an algorithm and applied that to large labeled databases, for instance LabelMe (Russell, 2008) allowed us to determine typical anchor objects as a function of different scene categories. To test whether the presence of such anchor objects has any behavioral relevance, we manipulated anchor objects using 3D rendered scenes (Boettcher et al., 2018). Participants in this study were then asked to search for various local objects with anchor objects either present or absent (replaced by a similarly sized and semantically consistent object). Tracking their eye movements we found prolonged search times and increased fixation distributions when anchors were absent, providing first evidence that the presence of anchors plays a crucial role in guiding efficient search in naturalistic scenes, and might prove to be also important for scene perception and object identification.

A better understanding of the role of anchor objects for the efficiency of object search could also improve the ability of models to predict eye movements during real-world search or improve identification of small or occluded objects. When looking for a mirror in an empty bathroom (see Fig. 5A), existing models like the contextual guidance model (Torralba et al., 2006) would have predicted a rather broad, horizontally unconstrained search area (see Fig. 5B). However, when we had participants in the lab search for an "invisible mirror", they mainly looked above the sink and nowhere else within that horizontal plane as can be seen in very confined fixation distributions (see Fig. 5C). A model that would include spatial predictions of objects based on anchors could greatly reduce predicted search areas by initially identifying and locating a scene's anchors, and restricting target search areas to locations predicted in relation to the according anchor, e.g. "above the sink" (see Fig. 5D). Such a model could make use of the fact that anchors tend to not only be big but also diagnostic of scene categories in that the strategy would be to identify and locat them first and only in a second step to identify and locate the search target itself. Humans appear to use such a multi-stage strategy. Machines might be able to learn from us.

We are only at the beginning of understanding the intricate ways in which predictions about objects in scenes are organized both in space and time. The role of object functions will need to be considered more explicitly in models trying to predict attention allocation in scenes since they have shown to influence attention allocation and perception (e.g., Castelhano & Witherspoon, 2016; Clement, O'Donnell, & Brockmole, 2019). Most indoor scenes were created to make every day actions more efficient: We tend to put our toothbrush and toothpaste close to each other and near a sink, because this will make it more efficient to brush our teeth than keeping the toothbrush in a distant closet. Thus, meaningful "phrases" within scenes tend to be established and organized by object functions and everyday actions or schemata (e.g. the "shower phrase" where you wash yourself, the "stove phrase" where you cook, the "desk phrase" where you work, etc.), a concept that has recently been further investigated by Josephs and Konkle (2019) "reachspaces". Such hierarchical structures of scenes and spatial arrangements of "phrases" or "reachspaces" most likely dramatically change when moving from indoor to outdoor scenes, and it remains to be seen whether natural, i.e. not manmade, scenes actually show such a hierarchical structure. The "functions" of nature might nevertheless show their own organizational grammar: In the mountains you know that close to a glacier you will be able to find moraines and gravel, and experienced mushroom seekers know how to "read" the weather and the forest in order to find their prey.

9. All things considered

Moving from highly abstract artificial tasks and simplified stimuli to more complex, diverse realities of the world brings about both novel, exciting questions and new hurdles to overcome. Stimulus control by necessity becomes more challenging, and the number of variables to consider becomes increasingly large and more difficult to measure. Finding the right balance between highly-controlled laboratory experiments while at the same time trying to move closer to what we consider the "real world" will be crucial in the years to come.

I have argued that understanding the meaning and structure of natural scenes is the key to understanding the efficiency of object and scene perception as well as search. Anchor objects seem to play a crucial role within the larger structure of scenes, predicting both locations and identities of other objects therein. We yet need to fully grasp (and compute) the intricate relationships between objects in scenes. Are they solely based on functions and tasks that we perform on a daily basis? Are anchor objects as we define them solely beneficial to object identification and search in manmade scenes, where humans have built their environment according to the needs of their daily routines?

Considering object functions, tasks and routines that we perform in our visual world, dynamics of actions and events become more important as well. Most of the research done on scene perception, however,

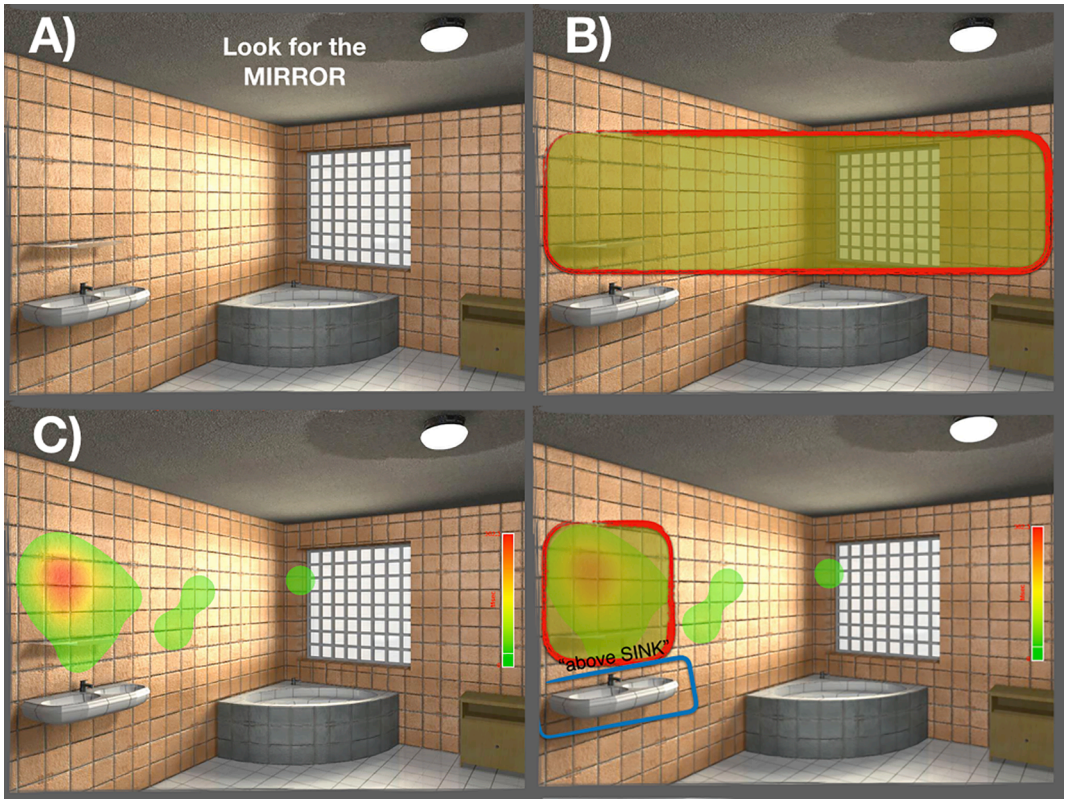

Fig. 5. Observers received the task to search for an "invisible mirror" in an empty bathroom containing only anchors (A). A model that only uses horizontally unrestricted contextual predictions would create a very large search space (B). A heat map based on fixation durations of participants searching for the invisible mirror (C). A model that would include anchor based predictions of objects could greatly reduce search space (D).

has dealt with still images. Scene grammar in the way we have conceptualized it does not explicitly include predictions about actions and events. Here the broader notion of schemata and scripts might be more applicable and worth trying to apply to some of the lingering questions in scene perception and search.

Finally, with the rise of deep neural networks (DNNs) one might start to wonder whether we are simply very good statisticians, who over the course over our lives have incorporated the meaning and structure of scenes and perfected the use of such statistics to efficiently process them; or, alternatively, there is more to the human mind, and if so, what? Recent developments of Generative Adversary Networks (GANs; for a review see Karras, Aila, Laine, & Lehtinen, 2018) have been able to produce scenes based on a latent grammar that they have learned. Does this type of grammar differ from ours? If so, how? Evident differences between how we and such algorithms process scene information include that humans gather experiences not merely through passive exposure, but by active interactions with the world. We move in 3D space, guided by curiosity, motivations, and emotions. Making use of DNNs, we can now directly compare their performance to ours, and by doing so I predict that we will remain fascinated by the machinery of the human mind.

10. Conclusions

Working in the field of scene perception means navigating at the intersection of many diverse fields like vision science, computer science, philosophy, linguistics, architecture, or design. While research over the past decades has provided us with fascinating insights regarding topics spanning from human ability for ultra rapid scene recognition and gist processing, over massive memory for objects and scenes to efficient attention allocation and eye movement control in scenes, many frontiers still need further exploration. More concerted efforts targeting the intersection of seemingly disparate fields of research could create new, powerful synergies, which might inform scene perception in hitherto unthinkable ways. I hope this review is able to spark new interest for scene perception enthusiasts as well as newcomers to the investigation of the meaning and structure of scenes.

CRediT authorship contribution statement

Melissa Le-Hoa Võ: Conceptualization, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

I would like to thank the Vision Sciences Society and Elsevier for honoring my work with the 2018 Young Investigator Award. I am grateful for the introduction to scene perception research by John Henderson as well as for the outstanding mentorship of Jeremy Wolfe, under whose guidance much of this work was accomplished. I also thank a large group of friends and family, colleagues and students who significantly contributed to the research described in this paper (Gioia, Cy, and Daniel Baldauf, Dejan Draschkow, Sage Boettcher, Tim Lauer, Tim Cornelissen, Sabine Öhlschläger, Laura Maffongelli, Jona Sassenhagen, as well as my beloved SGL gang past, present, and future). This work was funded by the Emmy Noether-Programm of the Deutsche Forschungsgemeinschaft (DFG) Grant VO 1683/2-1 awarded to MLV as

well as by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 222641018 – SFB/TRR 135, sub-project C7 to MLV.

References

Adeli, H., Vitu, F., & Zelinsky, G. J. (2016). A model of the superior colliculus predicts fixation locations during scene viewing and visual search. *Journal of Neuroscience*.

Baillargeon, R. (1987). Object permanence in 3½- and 4½-month-old infants. *Developmental Psychology, 23*(5), 655–664. https://doi.org/10.1037/0012-1649.23.5.655.

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

Becker, M. W., Pashler, H., & Lubin, J. (2007). Object-intrinsic oddities draw early saccades. *Journal of Experimental Psychology: Human Perception and Performance, 33*(1), 20–30. https://doi.org/10.1037/0096-1523.33.1.20.

Biederman, I. (1981). On the semantics of a glance at a scene.

Biederman. I. (1976). On processing information from a glance at a scene: some implications for a syntax and semantics of visual processing. UODICS'76. 75–88. https://doi.org/10.1145/1024273.1024283.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*(2), 143–177. https://doi.org/10.1016/0010-0285(82)90007-X.

Boettcher, S. E. P., Draschkow, D., Dienhart, E., & Võ, M.-H. (2018). Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search. *Journal of Vision, 18*(13), 11. https://doi.org/10.1167/18.13.11.

Bonitz, V. S., & Gordon, R. D. (2008). Attention to smoking-related and incongruous objects during scene viewing. *Acta Psychologica, 129*(2), 255–263. https://doi.org/10.1016/j.actpsy.2008.08.006.

Brockmole, J. R., & Le-Hoa Vo, M. (2010). Semantic memory for contextual regularities within and across scene categories: Evidence from eye movements. *Attention, Perception & Psychophysics, 72*(7), 1803–1813. https://doi.org/10.3758/APP.72.7.1803.

Brooks, B. M., Attree, E. A., Rose, F. D., Clifford, B. R., & Leadbetter, A. G. (1999). The specificity of memory enhancement during interaction with a virtual environment. *Memory, 7*(1), 65–78. https://doi.org/10.1080/741943713.

Castelhano, M. S., & Heaven, C. (2010). The relative contribution of scene context and target features to visual search in scenes. *Attention, Perception & Psychophysics, 72*(5), 1283–1297. https://doi.org/10.3758/APP.72.5.1283.

Castelhano, M. S., & Heaven, C. (2011). Scene context influences without scene gist: Eye movements guided by spatial associations in visual search. *Psychonomic Bulletin & Review, 18*(5), 890–896. https://doi.org/10.3758/s13423-011-0107-8.

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance, 33*(4), 753–763. https://doi.org/10.1037/0096-1523.33.4.753.

Castelhano, M. S., & Witherspoon, R. L. (2016). How you use it matters: Object function guides attention during visual search in scenes. *Psychological Science, 27*(5), 606–621. https://doi.org/10.1177/0956797616629130.

Castelhano, M., & Henderson, J. (2005). Incidental visual memory for objects in scenes. *Visual Cognition, 12*(6), 1017–1040. https://doi.org/10.1080/13506280444000634.

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology, 36*(1), 28–71. https://doi.org/10.1006/cogp.1998.0681.

Clement, A., O'Donnell, R. E., & Brockmole, J. R. (2019). The functional arrangement of objects biases gaze direction. *Psychonomic Bulletin & Review, 26*(4), 1266–1272. https://doi.org/10.3758/s13423-019-01607-8.

Coco, M. I., Nuthmann, A., & Dimigen, O. (2020). Fixation-related brain potentials during semantic integration of object–scene information. *Journal of Cognitive Neuroscience, 32*(4), 571–589. https://doi.org/10.1162/jocn_a_01504.

Cohn, N., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2014). The grammar of visual narrative: Neural evidence for constituent structure in sequential image comprehension. *Neuropsychologia, 64*, 63–70. https://doi.org/10.1016/j.neuropsychologia.2014.09.018.

Cornelissen, T. H. W., & Võ, M.-H. (2016). Stuck on semantics: Processing of irrelevant object-scene inconsistencies modulates ongoing gaze behavior. *Atten Percept Psychophys, 79*(1), 154–168. https://doi.org/10.3758/s13414-016-1203-7.

Cunningham, C. A., Yassa, M. A., & Egeth, H. E. (2015). Massive memory revisited: Limitations on storage capacity for object details in visual long-term memory. *Learning & Memory, 22*(11), 563–566. https://doi.org/10.1101/lm.039404.115.

de, P., Graef, Christiaens, D., & Gd'Ydewalle.. (1989). *Perceptual effects of scene context on object identification*. Psychological Research Psychologische Forschung.

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

Demiral, Ş. B., Malcolm, G. L., & Henderson, J. M. (2012). ERP correlates of spatially incongruent object identification during scene viewing: Contextual expectancy versus simultaneous processing. *Neuropsychologia, 50*(7), 1271–1285. https://doi.org/10.1016/j.neuropsychologia.2012.02.011.

Draschkow, D., & Võ, M.-H. (2016). Of "what" and "where" in a natural search task: Active object handling supports object location memory beyond the object's identity. *Attention, Perception & Psychophysics, 78*(6), 1574–1584. https://doi.org/10.3758/s13414-016-1111-x.

Draschkow, D., & Võ, M.-L.-H. (2017). Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific Reports, 7*(1), 16471. https://doi.org/10.1038/s41598-017-16739-x.

Draschkow, D., Heikel, E., Võ, M.-H., Fiebach, C. J., & Sassenhagen, J. (2018). No evidence from MVPA for different processes underlying the N300 and N400 incongruity effects in object-scene processing. *Neuropsychologia, 120*, 9–17. https://doi.org/10.1016/j.neuropsychologia.2018.09.016.

Draschkow, D., Reinecke, S., Cunningham, C. A., & Võ, M.-H. (2019). The lower bounds of massive memory: Investigating memory for object details after incidental encoding. *Quarterly Journal of Experimental Psychology, 72*(5), 1176–1182. https://doi.org/10.1177/1747021818783722.

Draschkow, D., Wolfe, J. M., & Vo, M. L. H. (2014). Seek and you shall remember: Scene semantics interact with visual search to build better memories. *Journal of Vision, 14*(8), 10. https://doi.org/10.1167/14.8.10.

Eckstein, M. P., Koehler, K., Welbourne, L. E., & Akbas, E. (2017). Humans, but not deep neural networks, often miss giant targets in scenes. *Current Biology, 27*(18), 2827–2832.e3. https://doi.org/10.1016/j.cub.2017.07.068.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General, 108*(3), 316–355. https://doi.org/10.1037/0096-3445.108.3.316.

Friedrich, M., & Friederici, A. D. (2006). Early N400 development and later language acquisition. *Psychophysiology, 43*(1), 1–12. https://doi.org/10.1111/j.1469-8986.2006.00381.x.

Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Cognitive Brain Research, 16*(2), 123–144. https://doi.org/10.1016/S0926-6410(02)00244-6.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Psychology Press.

Gilchrist, I. D., North, A., & Hood, B. (2001). Is Visual Search Really like Foraging? *Perception, 30*(12), 1459–1464. https://doi.org/10.1068/p3249.

Greene, M. R. (2013). Statistics of high-level scene context. *Frontiers in Psychology, 4*, 1–31. https://doi.org/10.3389/fpsyg.2013.00777.

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

Gronau, N., & Shachar, M. (2015). Contextual consistency facilitates long-term memory of perceptual detail in barely seen images. *Journal of Experimental Psychology: Human Perception and Performance, 41*(4), 1095–1111. https://doi.org/10.1037/xhp0000071.

Harman, K. L., Humphrey, G. K., & Goodale, M. A. (1999). Active manual control of object views facilitates visual recognition. *Current Biology, 9*(22), 1315–1318. https://doi.org/10.1016/S0960-9822(00)80053-6.

Hayhoe, M. M. (2016). Vision and action. *Annual Review of Vision Science, 3*(1), 389–413. https://doi.org/10.1146/annurev-vision-102016-061437.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*(4), 188–194. https://doi.org/10.1016/j.tics.2005.02.009.

Hayhoe, M. M., & Rothkopf, C. A. (2010). Vision in the natural world: Vision in the natural world. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(2), 158–166. https://doi.org/10.1002/wcs.113.

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision, 3*(1), 6. https://doi.org/10.1167/3.1.6.M3.

Helbing*, J., Draschkow*, D., & Võ, M. L.-H. (2020). Search superiority: Goal-directed attentional allocation creates more reliable incidental identity and location memory than explicit encoding in naturalistic virtual environments. *Cognition, 196*, Article 104147. https://doi.org/10.1016/j.cognition.2019.104147.

Helo, A., van Ommen, S., Pannasch, S., Danteny-Dordoigne, L., & Rämä, P. (2017). Influence of semantic consistency and perceptual features on visual attention during scene viewing in toddlers. *Infant Behavior and Development, 49*, 248–266. https://doi.org/10.1016/j.infbeh.2017.09.008.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour, 1*(10), 743–747. https://doi.org/10.1038/s41562-017-0208-0.

Henderson, J. M., & Hayes, T. R., Peacock, C., & Rehrig, G. (in press). Meaning and Attentional Guidance in Scenes: A Review of the Meaning Map Approach. Vision (Special Issue on Eye Movements and Visual Cognition).

Henderson, J. M., Brockmole, J. R., Castelhano, M., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes (pp. 1–26).

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review, 16*(5), 850–856. https://doi.org/10.3758/PBR.16.5.850.

Henderson, J. M., Weeks, P. A.,. J., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25*(1), 210–228. https://doi.org/10.1037//0096-1523.25.1.210.

Hespos, S. J., & vanMarle, K. (2011). Physics for infants: Characterizing the origins of knowledge about objects, substances, and number: Physics for infants. *Wiley Interdisciplinary Reviews: Cognitive Science, 3*(1), 19–27. https://doi.org/10.1002/wcs.157.

Holleman, G. A., Hooge, I. T. C., Kemner, C., & Hessels, R. S. (2020). The "Real-World Approach" and its problems: A critique of the term ecological validity. *Frontiers in Psychology, 11*, 721. https://doi.org/10.3389/fpsyg.2020.00721.

Hollingworth, A. (2006). Scene and position specificity in visual memory for objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(1), 58–69. https://doi.org/10.1037/0278-7393.32.1.58.

Hollingworth, A. (2012). Task specificity and the influence of memory on visual search: Comment on Võ and Wolfe (2012). *Journal of Experimental Psychology: Human Perception and Performance, 38*(6), 1596–1603. https://doi.org/10.1037/a0030237.
Henderson, A., & Henderson, J. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology General*.
Hopf, J.-M., Bader, M., Meng, M., & Bayer, J. (2003). Is human sentence parsing serial or parallel? *Cognitive Brain Research, 15*(2), 165–177. https://doi.org/10.1016/S0926-6410(02)00149-0.
Hwang, A. D., Wang, H.-C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research, 51*(10), 1192–1205. https://doi.org/10.1016/j.visres.2011.03.010.
Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*(10–12), 1489–1506. https://doi.org/10.1016/S0042- 6989(99)00163-7.
Johnson, S. P., Amso, D., & Slemmer, J. A. (2003). Development of object concepts in infancy: Evidence for early learning in an eye-tracking paradigm. *Proceedings of the National Academy of Sciences, 100*(18), 10568–10573. https://doi.org/10.1073/pnas.1630655100.
Josephs, E. L., Draschkow, D., Wolfe, J. M., & Võ, M.-H. (2016). Gist in time: Scene semantics and structure enhance recall of searched objects. *Acta Psychologica, 169*, 100–108. https://doi.org/10.1016/j.actpsy.2016.05.013.
Josephs, E., & Konkle, T. (2019). Perceptual dissociations among views of objects, scenes, and reachable spaces. *Journal of Experimental Psychology: Human Perception and Performance*, 1–26.
Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018, October 27). Progressive Growing of GANs for Improved Quality, Stability, and Variation. arXiv.org.
Kingstone, A., Smilek, D., & Eastwood, J. D. (2008). Cognitive Ethology: a new approach for studying human cognition. *British Journal of Psychology (London, England : 1953), 99*(Pt 3), 317–340. https://doi.org/10.1348/000712607X251243.
Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*, 219–227.
Koehler, K., & Eckstein, M. P. (2017). Beyond Scene Gist: Objects Guide Search More Than Scene Background. *Journal of Experimental Psychology: Human Perception and Performance, 1–18*. https://doi.org/10.1037/xhp0000363.
Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychological Science, 21*(11), 1551–1556. https://doi.org/10.1177/0956797610385359.
Kümmerer, Wallis, Gatys, & Bethge (2017). Understanding Low- and High-Level Contributions to Fixation Prediction. In The IEEE International Conference on Computer Vision (ICCV).
Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology, 62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123.
LaPointe, M. R. P., & Milliken, B. (2016). Semantically incongruent objects attract eye gaze when viewing scenes for change. *Visual Cognition, 24*(1), 63–77. https://doi.org/10.1080/13506285.2016.1185070.
Larson, A. M., & Loschky, L. C. (2009). The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision, 9*(10), 6. https://doi.org/10.1167/9.10.6.
Larson, A. M., Freeman, T. E., Ringer, R. V., & Loschky, L. C. (2014). The spatiotemporal dynamics of scene gist recognition. *Journal of Experimental Psychology: Human Perception and Performance, 40*(2), 471–487.
Lauer, T., Cornelissen, T. H. W., Draschkow, D., Willenbockel, V., & Võ, M. L. H. (2018). The role of scene summary statistics in object recognition. *Scientific Reports, 8*(1). https://doi.org/10.1038/s41598-018-32991-1.
Lauer, T., Willenbockel, V., Maffongelli, L., & Võ, M. L. H. (2020). The influence of scene and object orientation on the scene consistency effect. *Behavioural Brain Research, 394*, 112812. https://doi.org/10.1016/j.bbr.2020.112812.
Li, C.-L., Aivar, M. P., Kit, D. M., Tong, M. H., & Hayhoe, M. M. (2016). Memory and visual search in naturalistic 2D and 3D environments. *Journal of Vision, 16*(8), 9. https://doi.org/10.1167/16.8.9.
Loftus, G., & Mackworth, N. (1977). Cognitive Determinants of Fixation Location during Picture Viewing.
Mack, S. C., & Eckstein, M. P. (2011). Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *Journal of Vision, 11*(9), 9. https://doi.org/10.1167/11.9.9.
Maffongelli, L., Bartoli, E., Sammler, D., Kölsch, S., Campus, C., Olivier, E., Fadiga, L., & D'Ausilio, A. (2015). Distinct brain signatures of content and structure violation during action observation. *Neuropsychologia, 75*, 30–39. https://doi.org/10.1016/j.neuropsychologia.2015.05.020.
Maffongelli, L., Öhlschläger, S., & Võ, M. L.-H. (2020). The development of scene semantics: First ERP indications for the processing of semantic object-scene inconsistencies in 24-month-olds. *Collabra: Psychology, 6*(1), 17707.
Malcolm, G. L., & Shomstein, S. (2015). Object-based attention in real-world scenes. *Journal of Experimental Psychology General, 144*(2), 257–263. https://doi.org/10.1037/xge0000060.
Malcolm, G. L., Groen, I. I. A., & Baker, C. I. (2016). Making sense of real-world scenes. *Trends in Cognitive Sciences, 20*(11), 843–856. https://doi.org/10.1016/j.tics.2016.09.003.
Mudrik, L., Lamy, D., & Deouell, L. Y. (2010). ERP evidence for context congruity effects during simultaneous object–scene processing. *Neuropsychologia, 48*(2), 507–517. https://doi.org/10.1016/j.neuropsychologia.2009.10.011.
Mudrik, L., Shalgi, S., Lamy, D., & Deouell, L. Y. (2014). Synchronous contextual irregularities affect early scene processing: Replication and extension. *Neuropsychologia, 56*, 447–458. https://doi.org/10.1016/j.neuropsychologia.2014.02.020.
Neider, M. B., & Zelinsky, G. J. (2005). Scene context guides eye movements during visual search. *Vision Research, 46*(5), 614–621. https://doi.org/10.1016/j.visres.2005.08.025.
Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision, 10*(8), 20. https://doi.org/10.1167/10.8.20.
Nuthmann, A., Smith, T. J., Engbert, R., & Henderson, J. M. (2010). CRISP: A computational model of fixation durations in scene viewing. *Psychological Review, 117*(2), 382–405. https://doi.org/10.1037/a0018924.
Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology, 41*(2), 176–210. https://doi.org/10.1006/cogp.1999.0728.
Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences, 11*(12), 520–527. https://doi.org/10.1016/j.tics.2007.09.009.
Öhlschläger, S., & Võ, M.-H. (2016). SCEGRAM: An image database for semantic and syntactic inconsistencies in scenes. *Behavior Research Methods, 49*(5), 1780–1791. https://doi.org/10.3758/s13428-016-0820-3.
Öhlschläger, S., & Võ, M. L.-H. (2020). Development of scene knowledge: Evidence from explicit and implicit scene knowledge measures. *Journal of Experimental Child Psychology, 194*, Article 104782.
Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience, 6*(7), 674–681. https://doi.org/10.1038/nn1082.
Portilla, J., & Simoncelli, E. (n.d.). A parametric texture model based on joint statistics of complex wavelet coefficients. International Journal of Computer Vision.
Potter, M. C. (1975). Meaning in visual search. *Science, 187*(4180), 965–966.
Ratner, H. H., & Myers, N. A. (1981). Long-term memory and retrieval at ages 2, 3, 4. *Journal of Experimental Child Psychology, 31*(3), 365–386. https://doi.org/10.1016/0022-0965(81)90024-2.
Russell, & Torralba. (2008). LabelMe: a database and web-based tool for image annotation, 33.
Saarnio, D. A. (1990). Schematic knowledge and memory in young children. *International Journal of Behavioral Development, 13*(4), 431–446.
Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review, 99*(4), 605–632.
Stoll, J., Thrun, M., Nuthmann, A., & Einhäuser, W. (2015). Overt attention in natural scenes: Objects dominate features. *Vision Research, 107*, 36–48. https://doi.org/10.1016/j.visres.2014.11.006.
Tatler, B. W., Brockmole, J. R., & Carpenter, R. H. S. (2017). LATEST: A model of saccadic decisions in space and time. *Psychological Review, 124*(3), 267–300. https://doi.org/10.1037/rev0000054.
Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision, 11*(5), 5. https://doi.org/10.1167/11.5.5.
Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*(4), 766–786. https://doi.org/10.1037/0033-295X.113.4.766.
Trapp, S., & Bar, M. (2015). Prediction, context, and competition in visual recognition. *Annals of the New York Academy of Sciences, 1339*(1), 190–198. https://doi.org/10.1111/nyas.12680.
Trewartha, K. M., Case, S., & Flanagan, J. R. (2015). Integrating actions into object location memory: A benefit for active versus passive reaching movements. *Behavioural Brain Research, 279*, 234–239. https://doi.org/10.1016/j.bbr.2014.11.043.
Truman, A., & Mudrik, L. (2018). Are incongruent objects harder to identify? The functional significance of the N300 component. *Neuropsychologia, 117*, 222–232. https://doi.org/10.1016/j.neuropsychologia.2018.06.004.
Underwood, G., Humphreys, L., & Cross, E. (2009). Congruency, saliency and gist in the inspection of objects in natural scenes, 9.
Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? Evidence from eye movements during the inspection of real-world scenes. *Consciousness and Cognition, 17*(1), 159–170. https://doi.org/10.1016/j.concog.2006.11.008.
Vo, M. L. H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision, 9*(3), 24. https://doi.org/10.1167/9.3.24.
Võ, M. L. H., & Henderson, J. M. (2010). The time course of initial scene processing for eye movement guidance in natural scene search. *Journal of Vision, 10*(3). https://doi.org/10.1167/10.3.14, 14.1–13.
Võ, M. L. H., & Henderson, J. M. (2011). Object–scene inconsistencies do not capture gaze: Evidence from the flash-preview moving-window paradigm. *Atten Percept Psychophys, 73*(6), 1742–1753. https://doi.org/10.3758/s13414-011-0150-6.
Võ, M. L. H., & Wolfe, J. M. (2012). When does repeated search in scenes involve memory? Looking at versus looking for objects in scenes. *Journal of Experimental Psychology: Human Perception and Performance, 38*(1), 23–41. https://doi.org/10.1037/a0024147.
Võ, M. L. H., & Wolfe, J. M. (2013a). Differential electrophysiological signatures of semantic and syntactic scene processing. *Psychological Science, 24*(9), 1816–1823. https://doi.org/10.1177/0956797613476955.
Võ, M. L. H., & Wolfe, J. M. (2013b). The interplay of episodic and semantic memory in guiding repeated search in scenes. *Cognition, 126*(2), 198–212. https://doi.org/10.1016/j.cognition.2012.09.017.
Võ, M.-H., Boettcher, S. EP., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology, 29*, 205–210. https://doi.org/10.1016/j.copsyc.2019.03.009.

Võ, M. L. H., Schneider, W. X., & Matthias, E. (2008). Transsaccadic scene memory revisited: A "Theory of Visual Attention (TVA)"based approach to recognition memory and confidence for objects in naturalistic scenes. *Journal of Eye-Movement Research, 2(2):7*, 1–13.

Wolfe, J. M. (1998). Visual memory: What do you know about what you saw? *Current Biology, 8*(9), R303–R304. https://doi.org/10.1016/S0960-9822(98)70192-7.

Wolfe, J. M., Alvarez, G. A., Rosenholtz, R., Kuzmova, Y. I., & Sherman, A. M. (2011a). Visual search for arbitrary objects in real scenes. *Attention, Perception & Psychophysics, 73*(6), 1650–1671. https://doi.org/10.3758/s13414-011-0153-3.

Wolfe, J. M., Võ, M. L. H., Evans, K. K., & Greene, M. R. (2011b). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences, 15*(2), 77–84. https://doi.org/10.1016/j.tics.2010.12.001.

Wu, C. C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology, 5*, 1–14. https://doi.org/10.3389/fpsyg.2014.00054.