The Ingredients of Scenes that Affect Object Search and Perception

Tim Lauer and Melissa L.-H. Võ

1 Introduction

What determines where we attend and what we perceive in a visually rich environment? Since we typically cannot process everything that is in our field of view at once, certain information needs to be selected for further processing. Models of attentional control often distinguish two aspects: Bottom-up attention (sometimes referred to as "exogenous attention") focuses on stimulus characteristics that may stand out to us, while top-down (or "endogenous") attention focuses on goal-driven influences and knowledge of the observer (e.g., Henderson et al., 2009; Itti & Koch, 2001). In this chapter, we focus on top-down guidance of attention and object perception in scene context; particularly, on top-down guidance that is rooted in generic scene knowledge—or *scene grammar* as we will elaborate on later—and is abstracted away from specific encounters with a scene, but stored in long-term memory.

Suppose that you are looking for cutlery in a rented accommodation. You would probably search in the kitchen or in the living room but certainly not in the bathroom. Once in the kitchen, you would probably readily direct your attention to the cabinets—it would not be worthwhile to inspect the fridge or the oven. Despite having a specific goal, certain items may attract your attention, such as a bowl of fruits or colorful flowers on the kitchen counter. If you found forks, you might expect to find the knives close by. While viewing the kitchen, you would probably not have a hard time recognizing various kitchen utensils, even if they were visually small, occluded or otherwise difficult to identify. In this example, one benefits from context information, from prior experience with kitchens of all sorts. That is, in the real

T. Lauer (✉) · M. L.-H. Võ
Goethe University Frankfurt, Frankfurt am Main, Germany
e-mail: tlauer@psych.uni-frankfurt.de

B. Ionescu et al. (eds.), *Human Perception of Visual Information*,
https://doi.org/10.1007/978-3-030-81465-6_1

world, objects are hardly ever seen in isolation but typically in similar, repeating surroundings which allows us to make near-optimal predictions in perception and goal-directed behavior (Bar, 2004; Oliva & Torralba, 2007; Võ et al., 2019). Figure 1 provides an illustration: While it is difficult to recognize the isolated object in the left panel, the availability of scene context (right panel) probably helps in determining the identity of the object (here an electric water kettle).

In this chapter, we will first review how attention is allocated in the real world from a stimulus-driven perspective. We will then outline important aspects of attentional guidance during visual search, followed by a section on contextual influences on object recognition—an integral part of search. In particular, we focus on what types of contextual information or "ingredients" the visual system utilizes for object search and recognition, a question that has remained largely unexplored until recently. To this end, we refer to diverse methodologies (like psychophysics, eye tracking, neurophysiology, and computational modelling) used at different degrees of realism (ranging from on-screen experiments, via virtual reality to studies in the real world). Finally, we will bring the findings together, discussing the relative contributions of various context ingredients to object search and recognition, as well as future directions and mutual benefits of human and computer vision research.

2 Attentional Allocation in Real-World Scenes

2.1 The Role of Low-Level Features

The bowl of fruits in our introductory example (see Fig. 1) would be expected based on the semantic scene context, but might initially stand out to us in terms of low-level features (e.g., color) that differ from the surroundings (e.g., white kitchen

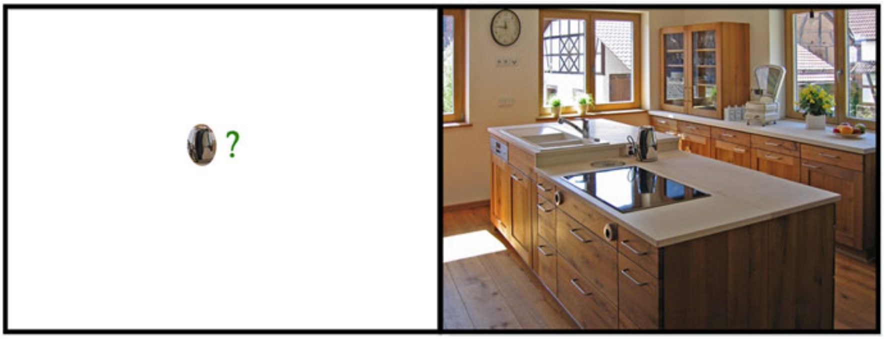


Fig. 1 While it is difficult to recognize the isolated object in the left panel, the kitchen context (right panel) may help in determining that the object is an electric water kettle. The kitchen scene was reproduced and adapted with permission from *Lignum Moebel*, Germany (https://lignum-moebel.de)

counter). Over the last two decades, several computational models of bottom-up, stimulus-driven attention have been put forth (for reviews, see Borji, 2019; Borji & Itti, 2013; Krasovskaya & Macinnes, 2019). A seminal early model of attention that inspired numerous other models is the saliency model by Itti and Koch (2000, 2001). Visual salience is defined as the "distinct subjective perceptual quality which makes some items in the world stand out from their neighbors and immediately grab our attention" (Itti, 2007). The model computes a salience map with regions that are likely attended by the observer based on low-level feature contrast (in intensity, orientation, and color) across spatial scales, motivated by receptive fields in the human visual system. Note that, as a proxy for *overt* visual attention, researchers often measure fixations and compare the empirical distributions to model predictions. However, visual attention is in principle not limited to the point of fixation and can be directed to regions outside of the fovea (commonly referred to as *covert* attention). Low-level saliency models have been shown to predict overt attention above chance under free viewing conditions (i.e., in the absence of a specific task), with highest predictability found for the first fixation (e.g., Parkhurst et al., 2002). Interestingly, these models capture where we direct our gaze merely based on low-level feature contrast, that is, without knowledge of image content or meaning (e.g., it is not known that the salient spot in the kitchen is a bowl of fruits or flowers).

2.2 The Role of Mid-Level Features and Objects

While low-level image features certainly play a decisive role for attentional allocation, it has been questioned whether attention is effectively attracted by such low-level features or rather higher-level features or objects that are not incorporated in low-level salience models (Einhäuser et al., 2008; Nuthmann & Henderson, 2010; Pajak & Nuthmann, 2013; Stoll et al., 2015). Objects often occur in locations that are salient (Spain & Perona, 2011)—oftentimes they make locations salient in the first place—and might thus be the driving force in attentional deployment (Schütt et al., 2019). Stoll et al. (2015) found that a state-of-the-art model of low-level salience and an object model predicted fixations equally well; however, when salience was reduced in regions that were relevant in terms of object content, the object model outperformed the salience model. Nuthmann and Einhäuser (2015) introduced a novel approach to investigate which image features influence gaze: Using mixed-effects models, they showed that mid-level features (e.g., edge density) and higher-level features (e.g., image clutter and segmentation) had a distinct contribution in gaze prediction as opposed to low-level features. Thus, many recent models incorporate mid to higher-level features in addition to low-level features to better predict fixation distributions in scene perception. To this end, deep neural networks (DNNs) have become increasingly popular and achieve benchmark performance in gaze prediction nowadays (Borji, 2019). One of the currently bestperforming networks, *DeepGazeII,* utilizes high-level features from a DNN trained on object recognition (Kümmerer et al., 2016).

2.3 The Role of Meaning

The role of scene meaning (or semantic informativeness) in attentional deployment while viewing real-world scenes has been studied for decades, and was recently systematically assessed by Henderson and colleagues (Henderson et al., 2018, 2019; Peacock et al., 2019a, 2019b). For a large number of local scene patches derived from scene images, they collected ratings of meaningfulness based on how informative or recognizable the patches were to observers. The authors then generated meaning maps which represent the spatial distribution of semantic features across a scene, comparable to a salience map (though not rooted in image-computable features). Meaning was shown to predict gaze successfully, as was low-level salience, but salience did not have a unique contribution when controlling for its correlation with meaning (Henderson & Hayes, 2017). This finding was replicated when predicting fixation durations instead of fixation distributions (Henderson & Hayes, 2018), and held across different tasks (Henderson et al., 2018; Rehrig et al., 2020), even when low-level image salience was highly task-relevant and meaning was not (Peacock et al., 2019a). However, it has been argued that the success of the meaning maps approach could be due to high-level image features that are not captured in classic salience models and could have strongly influenced observer's ratings of meaningfulness: *DeepGazeII,* which incorporates high-level object features, is able to outperform meaning maps at predicting fixations (Pedziwiatr et al., 2019).

Further, deriving meaning from objects in scenes has been shown to guide attention such that gaze tends to transition from one object to another object if the items are semantically related (Hwang et al., 2011; Wu et al., 2014a; for a review, see Wu et al., 2014b; see also De Groot et al., 2016). Objects that violate the global meaning of a scene (e.g., a mixer in the bathroom) strongly engage attention; they are typically looked at longer and more often than consistent objects (e.g., Cornelissen & Võ, 2017; De Graef et al., 1990; Friedman, 1979; Henderson et al., 1999; Loftus & Mackworth, 1978; Võ & Henderson, 2009b). While it has been established that attention can be "stuck" on these inconsistencies once they are spotted—even when they are irrelevant to one's current goals (Cornelissen & Võ, 2017, p.1)—it is a matter of debate whether they attract attention before they are fixated. Some studies have found semantic inconsistencies to influence initial eye-movements (e.g., the critical object is fixated earlier than a consistent object) (Becker et al., 2007; Bonitz & Gordon, 2008; Coco et al., 2019; Loftus & Mackworth, 1978; Nuthmann et al., 2019; Underwood et al., 2007, 2008; Underwood & Foulsham, 2006), yet other studies did not find indication for attention capture by inconsistencies (Cornelissen & Võ, 2017; De Graef et al., 1990; Furtak et al., 2020; Henderson et al., 1999; Võ & Henderson, 2009b, 2011). These mixed results may be related to characteristics of the scene stimuli (e.g., line drawings, photographs, or 3D-rendered scenes with varying degrees of clutter) and/or more or less controlled characteristics of the critical objects (e.g., size, eccentricity, salience).

With the rise of fully labeled image databases like LabelMe (Russell et al., 2008) assessing the semantic relatedness between objects and their scene contexts as well as inter-object relatedness has become easier. For instance, using graph theory by treating objects as nodes and assigning different weights to their connections has provided new avenues to determine clusters of semantically related objects within scenes—which we have started to call "phrases"—or prominent objects therein that anchor predictions about the location and identity of other objects nearby (for more details, see Sect. 4.3; Boettcher et al., 2018; for reviews, see Võ, 2021; Võ et al., 2019). Objects that do not fit their context tend to be regarded as surprising or interesting and can affect where we attend to in scenes.

2.4 The Role of Interestingness and Surprise

While the role of image features has been studied extensively (for reviews, see Borji, 2019; Borji & Itti, 2013; Krasovskaya & Macinnes, 2019), relatively little is known about how other factors such as interestingness or surprise modulate attentional deployment. Elazary and Itti (2008) proposed that interesting objects are in fact visually salient: Observers who contributed to the LabelMe database—a large collection of scenes with object annotations (Russell et al., 2008)—tended to label those objects that were salient even though they were free to choose which objects to label. In another study, when explicitly asked which scene locations are interesting, the choice of locations was largely similar across observers and correlated with fixation distributions of other observers (Masciocchi et al., 2009). Behavioral judgements and eye movements were also correlated with predictions of a salience model, yet not as highly as one would expect if salience was the only driving factor of interestingness. The authors concluded that there are both bottom-up and top-down influences on what we perceive as interesting and where we attend in an image (see also Borji et al., 2013; Onat et al., 2014). Other studies have shown that, beyond an influence of low-level salience, attentional allocation is modulated by the affective-motivational impact of objects or their importance for the scene ('t Hart et al., 2013; Schomaker et al., 2017), and that attention is attracted by surprising image locations in a Bayesian framework (e.g., Itti & Baldi, 2005). Moreover, some types of objects hold a special status: Text and faces, for instance, have been shown to greatly attract attention in scenes (see Wu et al., 2014b).

Taken together, inspired by early models of low-level salience, more recent research highlights the importance of higher-level features and indicates that attention in scenes is largely object-based—with some objects attracting and/or engaging attention more than others. While DNNs achieve benchmark performance in a variety of tasks nowadays and have become increasingly popular in fixation prediction, more research is needed to see how they will further our understanding of human attention mechanisms. Further, it will be crucial to shed more light on when during scene viewing various features exert influence on attentional allocation. Schütt et al. (2019) disentangled the contribution of low and higher-level features to fixation distributions over time, showing that the influence of low-level features is mostly limited to the first fixation and that higher-level features, as incorporated in *DeepGazeII*, predict fixations better starting 200 ms after stimulus onset. Despite the popularity of DNNs, a shortcoming of data-driven approaches is that they do not capture some aspects of human visual attention such as singleton (or “odd one out”) detection in artificial stimuli (even when the training data is adjusted, e.g., Kotseruba et al., 2020).

3 Guidance of Attention during Real-World Search

While the processing of image features can certainly play a role in where we attend, especially when free-viewing scenes, we are rarely ever mindlessly looking around. Instead, we tend to be driven by various agendas and task demands, one of which is the need to locate something or somebody. The interplay of bottom-up image features and more cognitively based, top-down influences during search is complex. As Henderson (2007) put it: “In a sense, we can think of fixation as either being “pulled” to a particular scene location by the visual properties at that location, or “pushed” to a particular location by cognitive factors related to what we know and what we are trying to accomplish” (p. 219). However, it should be noted that it is not always straightforward to strictly delineate between bottom-up and top-down influences (Awh et al., 2012; see also Teufel & Fletcher, 2020); we are certainly not claiming that the aspects presented here are one *or* the other.

Traditionally, visual search was studied using simple artificial displays of randomly arranged targets and distractors (e.g., “find the letter T among several instances of the letter L”). The main measure was—and still is— reaction time (RT) as a function of set size (i.e., the number of items in the display). With increasing set size, RT is consistently longer in such a task, in equal steps, indicating that attention is serially deployed to one item after another (see Wolfe, 2020; Wolfe & Horowitz, 2017). However, in some cases, it is not necessary to inspect all items in the display: In “classic guided search” theory, a limited set of target features (e.g., color, motion, orientation, size) can guide attention in a top-down manner, narrowing down the number of possible items (for reviews, see Wolfe, 2020; Wolfe et al., 2011b; Wolfe & Horowitz, 2017). For instance, when looking for a red “T” among some red and some black “L”s one can disregard all black items. To this end, “feature binding” takes place: The shape and the color of the target are bound together in order to reject distractors as well as recognize the target(s). While the field has learned a lot from these types of experiments that mostly used meaningless stimuli, search in real-world scenes seems to be strongly influenced by other guiding factors.

Scenes are not random assemblies of features but most often structured and meaningful, which allows us to perform searches with remarkable efficiency. For instance, when looking for a teddy in the bedroom, fixations tend to cluster around the bed even if the target is not present and cannot guide attention by means of its features (see Võ et al., 2019). Search for objects in scenes appears to be much more efficient than search for isolated objects in random arrays, although it can be challenging to define a scene's set size adequately (see Wolfe et al., 2011a). As proposed in the *cognitive relevance framework,* search in scenes is mainly guided by cognitive factors such as prior knowledge and current goals (Henderson et al., 2009; for a review, see Wolfe et al., 2011b).

What makes search in the real world so efficient despite the wealth and complexity of information contained in the visual input? While no one would doubt that scene context aids object search, relatively little is known about which "ingredients" of real-world scenes effectively guide attention, what their relative contributions are, and when they contribute during the search. In the following, we attempt to shed more light on these ingredients.

3.1 The Role of Scene Gist

One line of work addressed the question of whether an initial brief glance at a scene influences attentional allocation. Within a fraction of a second, observers can obtain the "gist" of a scene, a coarse representation of its spatial properties and meaning that does not require the selection of individual objects (Greene & Oliva, 2009a, 2009b; Rousselet et al., 2005). While there is no universal account of scene gist, many definitions (including ours), state that gist allows the categorization of scenes at a basic level. For instance, one may categorize a scene as a kitchen and tell that it comprises something like a kitchen counter but not yet grasp that there are a toaster and a mixer resting on any of the surfaces. That is, one may "see the forest without representing the trees" (Greene & Oliva, 2009a). A brief glance in the range of milliseconds is too short to make a saccade and thus to foveate selected parts of the scene in order to perceive them with fine detail. In fact, scene gist recognition does not depend on the high visual acuity of the fovea; it can be achieved even when the scene is blurred or when only peripheral information is available (e.g., Loschky et al., 2019). One fundamental aspect of scene gist is spatial layout information. As demonstrated in the *spatial envelope model* and supported by behavioral studies, scenes can be categorized based on their global properties, such as the global shape, without the need to identify any objects in the scene (Oliva & Torralba, 2001, 2006). This way of processing the scene is considered to be largely feed-forward and, in terms of search guidance, is assumed to take place on a "nonselective pathway" that parallels a "selective pathway" which binds features and recognizes individual objects (Wolfe et al., 2011b). Note that objects can also be an important source of information for scene categorization (MacEvoy & Epstein, 2011), especially for indoor scenes that are not always easily distinguishable in terms of their global properties.

To investigate how a brief glance at a scene guides search behavior, researchers have used the flash-preview moving window paradigm (Castelhano & Henderson, 2007; Võ & Henderson, 2010, 2011; Võ & Schneider, 2010; Võ & Wolfe, 2015): It initiates with a brief preview of a scene, followed by a target word and a search phase in which observers look for the target object in the original scene but through a gaze-contingent window that only reveals a small area of the scene at the current point of fixation. Given that the scene as a whole is not perceived during the search phase, this paradigm allows experimenters to assess the contribution of the scene's initial global percept to visual search. Note, however, that this contribution may be weaker under more natural search conditions in which the entire scene can be processed online during the search as well (see Võ & Wolfe, 2015). A scene's preview has been shown to influence visual search consistently in these studies, even when it was as short as 50 ms (Võ & Henderson, 2010). Võ and Schneider (2010) manipulated the type of context information that was available in the scene preview, selectively preserving either the global scene background or local objects (for an illustration, see Fig. 2). The availability of the scene background, conveying the spatial layout of the scene, resulted in faster detection of the targets and required fewer fixations compared to a control condition, whereas a preview of local objects did not facilitate search. Thus, a coarse representation of a scene's structure and meaning appears to already guide visual search effectively. Interestingly, knowing only the category of the scene does not seem to be sufficient, as was shown when a searched scene was

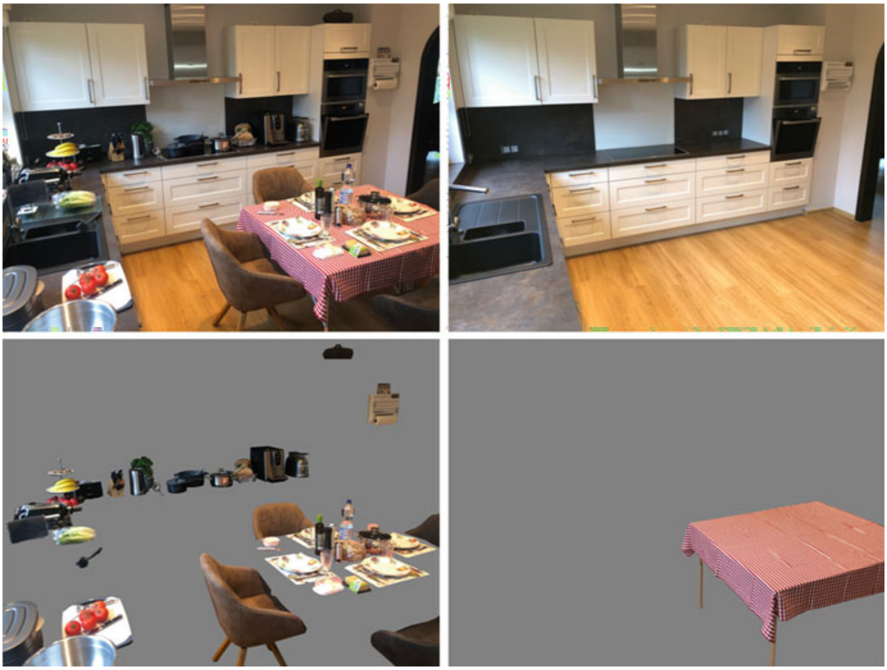


Fig. 2 Illustration of a kitchen scene (top left) that can be divided into the background (top right), local objects (bottom left) as well as an anchor object (bottom right)

primed by a different scene exemplar from the same category or by a word label of the category. Yet, a scene that is semantically inconsistent with a target (e.g., a mug of paint brushes in a bedroom) can facilitate search given that the object occurs in a reasonable location (Castelhano & Heaven, 2011, for a review see Castelhano & Krzyś, 2020).

The spatial layout of a scene can provide us with important constraints regarding the location of objects. For example, the occurrence of objects is constrained by the laws of physics such that objects rest on surfaces rather than hovering in the air. Even when we do not fully grasp a scene's meaning, we may be able to tell where its major surfaces lie (e.g., kitchen counters, tables, etc.) (see Fig. 2) and/or where the sky and the horizon are located. Moreover, two objects usually do not occupy the same physical space (Biederman et al., 1982), and we know where certain objects typically occur (e.g., a rug is often located on the floor) (Kaiser & Cichy, 2018; Neider & Zelinsky, 2006). Incorporating likely vertical object locations in a low-level salience model can significantly improve gaze prediction, as was demonstrated in the *contextual guidance model* (see Oliva & Torralba, 2006). More recently, the *surface guidance framework* was introduced, proposing that attention is allocated to surfaces in the scene that are related to the target object (Castelhano & Heaven, 2011; Pereira & Castelhano, 2014, 2019; for a review, see Castelhano & Krzyś, 2020).

3.2 The Role of Local Objects

Another line of work investigated the influence that selected parts of the scene, specifically objects, have on attentional allocation. In a naturalistic search task, Mack and Eckstein (2011) instructed participants to search for objects on tables while wearing mobile eye tracking glasses. The target object (e.g., a fork) was either located near a so-called cue object with which it would likely co-occur in natural scenes (e.g., a plate) or elsewhere (close to other objects). Targets were found faster if they were located near cue objects, and cue objects were fixated more frequently than other objects surrounding the targets, suggesting that object co-occurrence in the real world can boost search performance. In another study, in which participants inspected scene images or searched for targets therein, the LabelMe database of scenes with object annotations was used to determine the semantic relatedness of the currently fixated object to other objects in the scene or to the search target (Hwang et al., 2011). Gaze was shown to transition more likely to objects that are semantically related to the currently fixated object, even when the objects were not in close proximity. Moreover, the search data revealed that the influence of target-based semantic guidance increased throughout the trial. The finding of likely transitioning between related objects was replicated even when the objects were cropped (removed) from the scenes but not when discarding spatial dependencies among the cropped objects by re-arranging them (Wu et al., 2014a). When a preview of the original scene was added in order to provide gist information, there was no indication of increased semantic guidance. Moreover, there is evidence that the functional arrangement of objects influences gaze direction in the absence of scene context (e.g., a key that is arranged such that it can or cannot be inserted in a lock) (Clement et al., 2019). In object arrays, semantic information can be extracted extrafoveally and can guide even the first eye movement during search (Nuthmann et al., 2019). Taken together, both the semantic relation of objects as well as their spatial dependencies appear to be relevant for attentional allocation during search.

3.3 The Role of Anchor Objects

There seem to be certain objects that predict not only the occurrence, but particularly the location of other objects within a scene. Boettcher et al. (2018) explored the role of spatial predictions in object-based search guidance, introducing the concept of *anchor* objects. Anchors are typically large, static objects (i.e., they are rarely moved) that give rise to strong predictions regarding the identity and location of *local* objects clustering around them (e.g., the table may predict the position of a chair, a glass of water, and the salt). By contrast, local objects do not necessarily predict the location of other local objects (e.g., when searching for the salt, the location of a glass might not be that informative) (see Fig. 2). Using the LabelMe database, the concept of anchor objects was operationalized through four factors: variance of spatial location, frequency of co-occurrence, object-to-object distance, and clustering of objects (see Boettcher et al., 2018; c.f. Võ et al., 2019). In a series of eye tracking experiments, observers searched for target objects in images of 3D-rendered scenes (e.g., bathroom) that were manipulated to either contain a target-relevant anchor (e.g., shower) or a substitute object that was chosen to also be semantically consistent with the scene and of similar size (e.g., cabinet). Compared to the substitute objects, relevant anchors affected search performance such that there was a reduction in reaction time, scene coverage, and the time to transition from the anchor to the target. In line with this, in a recent virtual reality experiment, participants were slower at locating target objects when anchors were concealed by grey cuboids of similar dimensions compared to when they were fully visible (Helbing et al., 2020). Randomly re-arranging the anchors (or cuboids) resulted in an opposite effect, that is, targets were located faster in the cuboid condition, suggesting that both the identity and spatial predictions of anchors are crucial for their ability to guide search. Note that these inherent spatial predictions distinguish anchor objects from the notion of diagnostic objects (e.g., MacEvoy & Epstein, 2011) which may be important for conveying scene meaning and facilitating scene categorization, but need not yield precise predictions of the occurrence of other objects (Võ et al., 2019). It seems likely that anchor objects can be identified even in the periphery (see Koehler & Eckstein, 2017b, for a demonstration of peripheral extraction of object cues) and thus they might provide an effective way to locate smaller targets, building a bridge between the global scene and local objects.

3.4 Scene Grammar

Taken together, while scene gist yields an initial coarse representation of the scene's structure and meaning that can already effectively narrow down the search space, selected objects allow for a more fine-grained type of guidance (see also Wolfe et al., 2011b). Recent studies on the role of anchor objects showed that objects are not all equal in their ability to guide search. Rather, scenes appear to be hierarchically organized, with anchors being the core of so-called "phrases" that constitute meaningful subunits within a scene (e.g., the "shower phrase" versus the "toilet phrase"). Within these phrases, anchors hold stronger predictions about other local objects therein (e.g., the shampoo is *in* the shower and the toilet brush *next to* the toilet). When searching for the toilet paper in a bathroom, one can substantially reduce search time (and stress!) by outright avoiding to search the non-relevant shower and sink phrases.

While some of the regularities inherent in scenes have already been described decades ago (e.g., Biederman et al., 1982; Boyce & Pollatsek, 1992; Palmer, 1975), they are nowadays directly measurable using large-scale annotated databases and descriptive statistics (Greene, 2013, 2016; Russell et al., 2008). In analogy to the language domain, we have been referring to implicitly acquired knowledge of various regularities in scenes regarding *what* objects tend to be *where* as *scene grammar* (for reviews, see Võ, 2021; Võ et al., 2019; Võ & Wolfe, 2015). In language, semantics refers to conceptual relations between words while syntax describes the rules of sentence structure. Accordingly, we have used the terms scene semantics and syntax to describe the meaningfulness of object-scene relations (e.g., a pot belongs in the kitchen, not in the bathroom) or structural nature of these relations (the pot belongs on top of the stove, not on the floor), respectively (Võ & Henderson, 2009b). Violations of scene grammar have been shown to impede search performance and strongly influence eye-movements (e.g., Võ & Henderson, 2011). For instance, both semantic and syntactic violations are typically fixated longer and more often than their consistent counterparts (e.g., Võ & Henderson, 2009a). When objects are positioned inconsistently in a scene, it also takes longer to decide whether an object is the target or not once it is fixated (e.g., Võ & Wolfe, 2013b). Thus, contextual regularities may not only affect search guidance but also object recognition at various stages of the search.

4 Object Recognition in Scene Context

Object recognition is an integral part of search: Distractors need to be evaluated as to whether they are target candidates or not, and eventually the target needs to be identified and matched against the search template (for more details, see Sect. 5). In the following, we will outline how scene context affects object perception. Contextual influences on object perception were studied extensively for decades—

though mostly isolated from visual search—using behavioral measures, and more recently also neurophysiological methods. One of the core questions of this line of work has been at what stage(s) of object processing contextual modulation occurs.

4.1 Behavioral Work

Traditionally, the influence of scene context on object processing was studied using line drawings (Biederman et al., 1982; Boyce et al., 1989; Boyce & Pollatsek, 1992; Hollingworth & Henderson, 1998, 1999; Palmer, 1975). In Biederman's et al. (1982) influential object detection paradigm, observers were shown a target word (e.g., fire hydrant), followed by a briefly presented line drawing of a scene and a pattern mask with a location cue. Observers were asked if they had or had not seen the target object in the cued location. Target objects that were consistent with the scene context (e.g., a fire hydrant on the street) were detected faster and more accurately than semantically or syntactically inconsistent objects (e.g., a fire hydrant in the kitchen or a fire hydrant positioned in the air above a street, respectively) (see also Boyce et al., 1989), or other forms of violations (objects in unlikely rather than impossible locations, or objects with abnormal sizes). However, the consistency advantage was not replicated when taking response bias into account (Hollingworth & Henderson, 1998, 1999), which lent support to the *functional isolation model*, proposing that there is no interaction of scene and object processing on a perceptual level.

More recently, researchers have used color or grayscale photographs of scenes and an object naming task to probe the role of context in object recognition (Davenport & Potter, 2004; Lauer et al., 2018; Lauer et al., 2020a; Munneke et al., 2013; Sastyin et al., 2015). Observers were briefly presented with a scene containing a consistent or inconsistent object cutout in the foreground (or an isolated object superimposed on a scene; Lauer et al., 2018, 2020a, 2020b), followed by a perceptual mask and a response window, where they typed in the name of the object. In this paradigm, which is not prone to response bias and overcomes some limitations of early behavioral work (see Davenport & Potter, 2004), consistent objects were named more accurately than inconsistent objects across studies. Here, we refer to this effect as *scene-to-object consistency effect.* Moreover, scenes are named more accurately if they contain a consistent versus inconsistent object in the foreground (*object-to-scene consistency effect*), suggesting that objects and scenes are processed interactively (Davenport & Potter, 2004; see also Davenport, 2007; Leroy et al., 2020). The scene-to-object consistency effect cannot be explained by mere low-level feature overlap between the context and the target, nor does it depend on overt attention being directed to the object (Leroy et al., 2020; Munneke et al., 2013). Interestingly, the magnitude of the scene-to-object consistency effect is modulated by viewpoint: Objects that are seen from a canonical (easy) angle evoke a weaker effect than objects seen from a non-canonical (difficult) angle (Sastyin et al., 2015). Contextual modulation also depends on the displayed size of the target, with stronger effects seen for smaller objects that are more difficult to interpret (Zhang et al., 2020).

In the *model of contextual facilitation* (Bar, 2004), the gist of a scene rapidly activates scene schemata and yields predictions of associated objects that are matched against incoming information of the target, boosting its identification. We have recently probed if the scene-to-object consistency effect reflects such facilitation of object processing on a perceptual level by contrasting accuracy for isolated objects superimposed on scenes with accuracy for objects on unrecognizable scrambled scenes (baseline) (Lauer et al., 2020a). Consistent objects on scenes were named more accurately than consistent objects on scrambled scenes, suggesting that scene context indeed facilitated object recognition. Moreover, inconsistent objects on scenes were named less accurately than inconsistent objects on scrambled scenes, suggesting that the consistency manipulation also interfered with performance, either on a perceptual stage (e.g., by yielding misleading predictions) or a post-perceptual stage (e.g., through a mismatch detection that interfered with performance). It should be noted, however, that some other studies did not find facilitation of object recognition when the scene context was present versus absent (Davenport & Potter, 2004; Lauer et al., 2018; Roux-Sibilon et al., 2019). Possibly, facilitation effects are not always robust in the case of salient foreground objects or isolated objects for which figure-ground segmentation is arguably easy. In some paradigms, a strong segmentation advantage in the baseline (no-context condition) may also contribute to the absence of facilitation effects (see Davenport & Potter, 2004). Under more natural conditions—when objects are embedded in scenes and segmentation demands are also present in the baseline (e.g., by providing minimal context around the target)—facilitation was repeatedly shown, particularly pronounced for smaller targets that are more difficult to interpret (for extensive demonstrations, see Zhang et al., 2020; see also Brandman & Peelen, 2017). Besides distinct segmentation demands in the case of embedded objects, these objects also differ from isolated objects such that the context can yield spatial predictions and estimates of the target's size, potentially increasing the magnitude of contextual facilitation of object recognition.

In another recent study, objects were either presented within a scene or outside of it on the same horizontal or vertical plane (either unilateral or bilateral) (Leroy et al., 2020). Across manipulations, consistent objects and consistent scenes were named more accurately than inconsistent objects and scenes, respectively, confirming the reciprocal nature of the object-scene consistency effect. Given that contextual modulation was robust even when objects and scenes were not embedded in the same percept, these findings also suggest that, rather than arising at the earliest stages of object processing, scene context effects may occur at the stage of matching visual information with prior knowledge.

4.2 Neurophysiological Work

Context effects on object processing have also been investigated in neurophysiological studies, using electroencephalography. Specifically, event-related potentials (ERPs) provide a temporally precise measure that can track context effects online during stimulus exposure. The most widely studied context-sensitive ERP component is the N400—a negative deflection that peaks about 400 ms post stimulus onset—which was originally reported in the language domain: Sentences with a semantically inconsistent versus consistent word typically evoke a centrally distributed N400 effect, suggesting an impedance of semantic access (Kutas & Hillyard, 1980, 1983; for a review, see Kutas & Federmeier, 2011). In the scene perception domain, consistent versus inconsistent objects in (or superimposed on) scenes have been shown to evoke an N400 response with a comparable time course and topography (Draschkow et al., 2018; Ganis & Kutas, 2003; Lauer et al., 2018; Lauer et al., 2020a; Mudrik et al., 2010, 2014; Truman & Mudrik, 2018; Võ & Wolfe, 2013a; Zucker & Mudrik, 2019), indicating impeded access or integration of an object in a semantically inconsistent scene context (Mudrik et al., 2010, 2014). Moreover, across those studies that used scene stimuli, semantic consistency manipulations evoked an earlier negativity with a sometimes more frontal maximum known as N300 (but see Ganis & Kutas, 2003). This component has been suggested to reflect context effects on a more perceptual level, before object identification is completed (e.g., Mudrik et al., 2010, 2014). Specifically, it may reflect the difficulty of matching incoming information of the target with (misleading) predictions yielded by the inconsistent scene context. While it has been established that scene context can modulate object processing before object identification is completed (Lauer et al., 2020; Leroy et al., 2020; Truman & Mudrik, 2018; see also Brandman & Peelen, 2017), it is still debated whether the N300/N400 components are actually distinguishable in terms of the underlying processes or not: In a recent study from our laboratory, the two components were found to widely share neuronal activity patterns in a time-generalized decoding analysis (Draschkow et al., 2018). As opposed to semantic violations in scenes, syntactic violations (e.g., a towel on the bathroom floor) do not evoke N300/N400 effects but a later positivity that is comparable to the P600 frequently reported for grammatical violations in language (Võ & Wolfe, 2013a). A differential response to structural, "syntactic" inconsistencies has also been found in comic strips (Cohn et al., 2014) and action sequences (Maffongelli et al., 2015). Thus, the brain seems to distinguish the processing of object-scene relations in terms of their meaning and structural nature.

Where is scene meaning processed in the brain, and where does it influence object perception? In a functional magnetic resonance imaging (fMRI) experiment, Brandman and Peelen (2017) presented observers with isolated degraded (i.e. pixelated) objects, degraded objects in scenes, or scenes without any target objects, and found indication of contextual facilitation in the visual cortex (specifically in regions lateral occipital and posterior fusiform sulcus): Decoding accuracy for degraded objects in scenes exceeded accuracies for the other two conditions in a supra-additive manner, that is, it was greater than the sum of accuracies for these conditions. Interestingly, the effect of contextual facilitation was correlated with activity in regions that are crucial for scene processing (e.g., the parahippocampal place area and the retrosplenial cortex) (for a review on scene-selective regions and their functions, see Epstein & Baker, 2019). Magnetencephalography (MEG) data revealed that supra-additive facilitation emerged around 320 ms post stimulus onset, which is relatively late compared to a feedforward type of object processing in the absence of scene context influences (Cichy et al., 2014). It should be noted that the magnitude of contextual facilitation appears to depend on the visual characteristics of the target: On the behavioral level, facilitation of object detection was correlated with object ambiguity, with reduced facilitation seen for easier to identify, intact objects. In line with previous studies, these findings suggest that contextual modulation can arise on a perceptual level, and point to separate scene and object processing pathways that may interact in the visual cortex. Another recent study complemented these findings by showing signs of contextual facilitation in scene-selective areas when presenting degraded scenes with intact objects (Brandman & Peelen, 2019). Besides a reciprocal type of object-scene facilitation, there is also evidence of multimodal facilitation of object processing through auditory and semantic cues (Brandman et al., 2019).

Taken together, recent behavioral and neurophysiological work suggests that scene and object processing are not functionally isolated but that there are reciprocal influences facilitating perception, especially when the target stimulus is difficult to interpret.

4.3 Which Scene Ingredients Affect Object Processing?

Over the last few decades, numerous studies have demonstrated that scene context influences object perception, however, it was hardly ever asked which context ingredients the visual system actually utilizes, and at which time points they are relevant. The few studies that have probed individual scene properties are outlined below, grouped as studies employing *global* or *local* manipulations (affecting the context as a whole or parts of it, respectively).

Global influences on object processing. In a behavioral study, Brady et al. (2017) briefly presented observers with an object primed by either a grayscale scene or a texturized scene with a similar spatial distribution of orientations and spatial frequencies, preserving the global shape of the scene but no recognizable objects (Oliva & Torralba, 2006) (see Fig. 3). Objects primed by a semantically consistent scene were named more accurately than objects primed by an inconsistent scene. Critically, a similar but weaker scene-to-object consistency effect was found for texturized scenes, indicating that global scene properties—specifically spatial layout information—can modulate object recognition even in the absence of semantic object information. This finding is in line with studies highlighting the importance of

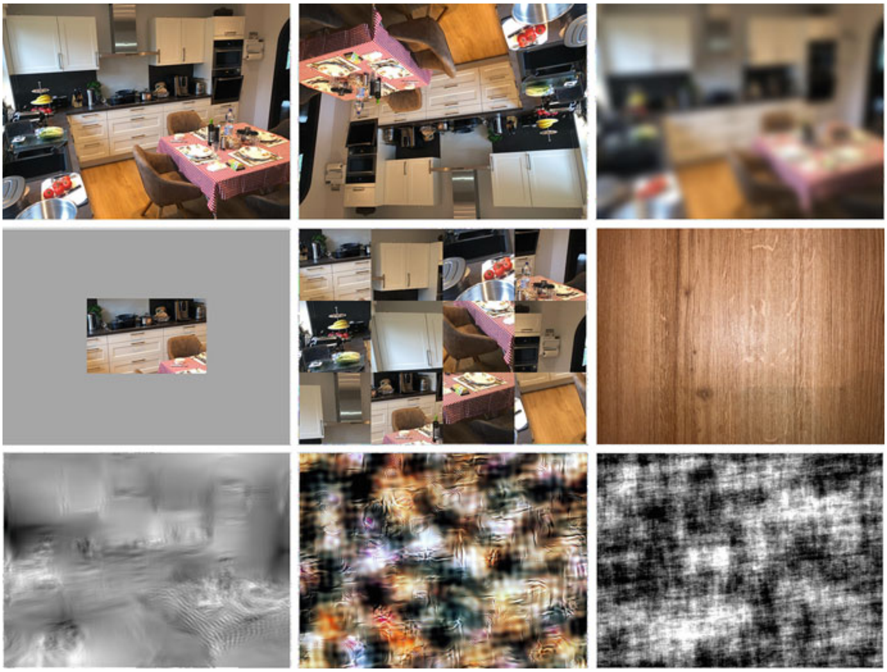


Fig. 3 Illustration of global context manipulations. From top left to bottom right: original scene, inverted scene, blurred scene, context area, jigsaw 4 × 4, material, layout texture (Brady et al., 2017), scene texture (Lauer et al., 2018), phase-scrambled scene. Note that the images only serve for illustration purposes; they are not a reproduction of the stimuli and parameters used in original studies

global scene properties for rapid scene understanding and categorization (Greene & Oliva, 2009a, 2009b; Joubert et al., 2007; Oliva & Torralba, 2001, 2006; Rousselet et al., 2005).

In a related study from our group, we presented consistent and inconsistent thumbnail objects superimposed on colored scenes, scene textures, or scrambled scenes (color controls) (Lauer et al., 2018). Our way of texturizing the scenes was different such that we preserved global scene summary statistics (including first to second order statistics as well as magnitude and phase correlation; see Portilla & Simoncelli, 2000) while discarding object semantics *and* spatial layout information (see Fig. 3). For scenes, we found a consistency effect at the behavioral level as well as an N300/N400 effect in ERPs. For textures, we found a non-significant trend in the same direction at the behavioral level as well as a significant N300/N400 response with a comparable time course, though less pronounced. Scrambled scenes, that retained color characteristics of the original scenes, did not show such effects. Thus, low-level scene statistics, as preserved in the textures, may modulate object processing even in the absence of spatial layout information while mere color information appears to be insufficient. It should be noted, however, that there was no indication of facilitation at the behavioral level, neither for scenes nor for textures, suggesting that the context effects may have been driven by interference in this study. By contrast, Zhang et al. (2020) found strong facilitation for objects embedded in scenes but still no facilitation for object cutouts on textures: Accuracy was even slightly higher in the baseline (minimal texture context) compared to a condition with full-size texture context. Note that there was no consistency manipulation in this study. A computer vision model (for more detail, see Sect. 5.2) was able to achieve higher accuracy for objects on textures than in the baseline, yet only when the target objects were small, and the facilitation effect was considerably weaker than the one found for original scenes.

Moreover, recent unpublished data from our laboratory suggests that objects in the context of global material backgrounds (e.g., a chair on wood vs. chair on water) (see Fig. 3) but not in the context of scrambled materials (color controls) yield a marginal consistency effect on the behavioral level as well as N300/N400 responses that are comparable to those found for scenes albeit weaker. In another study, we explored the role of object and scene orientation in the scene consistency effect (Lauer et al., 2020a). Specifically, we used inversion, a global manipulation that preserves low-level image properties (except for phase) but may interfere with semantic processing (see Fig. 3). Behaviorally and in ERPs, we found indication that upright scenes modulate the processing of both upright and inverted objects but that inverted scenes only modulate the processing of inverted objects. Corroborated by a later occurrence of ERP effects for inverted versus upright scenes, we argued that scene inversion may interfere with rapid scene gist-recognition, resulting in a later emergence of contextual influences on object processing.

Further, the amount of visual context that is available to the observer (quantified as revealed image area without the target divided by the target's size) was shown to strongly influence object recognition performance (illustrated in Fig. 3); facilitation was particularly strong in the case of small, difficult to perceive targets (Zhang et al., 2020). Moreover, contextual modulation by full scenes was robust even when the image was moderately blurred (Gaussian with $M = 0, SD \leq 8$, image size = 1024 × 1280 pixels) (see Fig. 3 for an illustration) but not when it was blurred more strongly ($SD > 8$, see Zhang et al., 2020), suggesting that context effects do not depend on fine detail conveyed by high spatial frequencies. In addition, the role of global spatial configuration of scene parts was investigated. To this end, full scenes were divided in equal parts (2 × 2, 4 × 4 or 8 × 8 "jigsaw") that were randomly re-arranged while the part that contained the target remained in its original position (see Fig. 3). Intact configuration facilitated object recognition compared to a minimal context (control) condition. There was also an increase in accuracy compared to the inconsistent 4 × 4 and 8 × 8 configurations; however, the inconsistent 2 × 2 configuration resulted in a similar performance, possibly indicating that large scene parts already convey sufficient context information even when the global configuration of the scene is inconsistent.

The role of peripheral vision in foveal object recognition was explored by Roux-Sibilon et al. (2019). Objects surrounded by a semantically consistent peripheral scene context (beyond 6 or 8 degrees) were categorized faster than those surrounded by an inconsistent context when there was a preview of the peripheral scene. Moreover, altering the phase coherence of the targets resulted in a lower visibility threshold for consistent objects than for inconsistent ones. The context effects were not observed in the case of phase-scrambled peripheral scenes, maintaining the power spectrum but no scene summary statistics or shape information (see Brady et al., 2017; Lauer et al., 2018).

Taken together, these studies indicate that there are global influences of context on object processing which do not depend on selected parts of the context but rather on a coarse representation of context as a whole. In the following, we will explore the role of more local influences on object processing.

Local influences on object processing. One type of local information that may be relevant for object processing is the presence of other objects. A number of studies have manipulated semantic object-to-object relation in the absence of scene context, for example, by priming a target object with a related or unrelated object, or by simultaneously presenting a target with surrounding object(s). An influence of relatedness was frequently found on the behavioral level (e.g., Auckland et al., 2007; Henderson et al., 1987) as well as on the neuronal level (e.g., Barrett & Rugg, 1990; Kovalenko et al., 2012; Li et al., 2019; McPherson & Holcomb, 1999). Besides semantic relatedness, spatial object-to-object relation has been shown to modulate object processing. For instance, two related objects are named more accurately or classified faster if their spatial arrangement is typical compared to atypical (e.g., a lamp on a desk vs. a lamp under a desk, respectively), given that both objects are attended (Gronau & Shachar, 2014; Roberts & Humphreys, 2011; see also Gronau, 2020). Moreover, there is electrophysiological evidence that semantic relatedness and spatial relation interacts in object processing (Quek & Peelen, 2020). However, the role of these object-to-object effects in the presence of scene context—in the presence of other visual information—is yet to be explored.

Only two behavioral studies, to our knowledge, have jointly investigated the influence of scene background and objects. One study found an influence of scene background on object detection but no influence of relatedness among the (five) objects in the scene (Boyce et al., 1989). However, the absence of a local context effect in this study may have been due to characteristics of the stimuli (e.g., line drawings of scenes with small objects), as pointed out by Davenport (2007). In a more recent object naming experiment, observers were presented with two foreground objects, either related or unrelated, in a scene that was either consistent with both objects, one object, or neither. Scene-to-object consistency resulted in higher accuracy, as did object-to-object relatedness, without interaction of the two variables (Davenport, 2007). In a related study from our laboratory, we explored the temporal dynamics of these types of context effects using EEG (Lauer et al., 2020b). We only found N300/N400 ERP responses when both objects were unrelated and inconsistent with the scene in comparison to all other conditions; all other possible comparisons were not significant, indicating that one congurent relation of an object with either the scene or the neighboring object is sufficient to eliminate the N300/N400 inconsistency effect in this type of paradigm. Thus, we found some indication of both global and local context effects, with no apparent difference in the timeline, in accordance with an interactive view of scene perception (Davenport & Potter, 2004). It should be noted that in these studies, the background scenes contained objects that may have contributed to the context effects. Moreover, in our study, the critical objects were salient and close to the point of fixation. Future studies might want to assess the influence of other local scene properties and viewing conditions.

5 Concluding Remarks and Future Directions

Although visual search and object recognition are typically two distinct research areas with varying experimental setups, in the following, we attempt to bring the findings outlined in this chapter together, pointing out some similarities and apparent differences between the two domains. Further, we will discuss the question of relative contributions of context ingredients, and conclude with a section on reciprocal benefits of human and computer vision research.

To begin with, visual search usually includes object recognition at various stages. For instance, every distractor needs to be evaluated as to whether it is a target candidate or not. Once the target is foveated this critical object needs to be identified and matched against the search template. Thus, benefits of scene context are not only due to more efficient guidance by scene grammar, but likely also due to improved object recognition leading to faster disengagement of distractors and target identification—the latter is usually measured as decision time (i.e., the time from initial target fixation to button press indicating the termination of search). Figure 4 provides an illustration of a search for a toaster in the kitchen.

Accordingly, evidence from both literatures suggests that scene gist, which can be inferred from the spatial layout of a scene (c.f. Võ & Wolfe, 2015), can readily modulate both search performance and object processing. This type of context information is available very rapidly and allows narrowing down search space (e.g., we would search for a toaster on that large horizontal surface in what appears to be a kitchen) or the number of possible object identities (e.g., the item is probably an electronic device, not a rock), respectively. In fact, even 50 ms of exposure to context is sufficient to affect search (Võ & Henderson, 2010) as well as object recognition (Zhang et al., 2020)—this number can even be lower in the absence of backward masking (e.g., 25 ms in the case of object recognition, Zhang et al., 2020).

Moreover, it has been established that local properties of a scene, specifically co-occurring objects, are an important source of information for target localization as well as identification. To this end, both the semantic relatedness of co-occurring local objects as well as their spatial dependencies are utilized. In a recent virtual reality study from our laboratory, anchor objects in scenes not only guided search but also significantly reduced the decision time once the target object was fixated, compared to a condition in which the anchors were concealed by gray cuboids (Helbing et al., 2020).

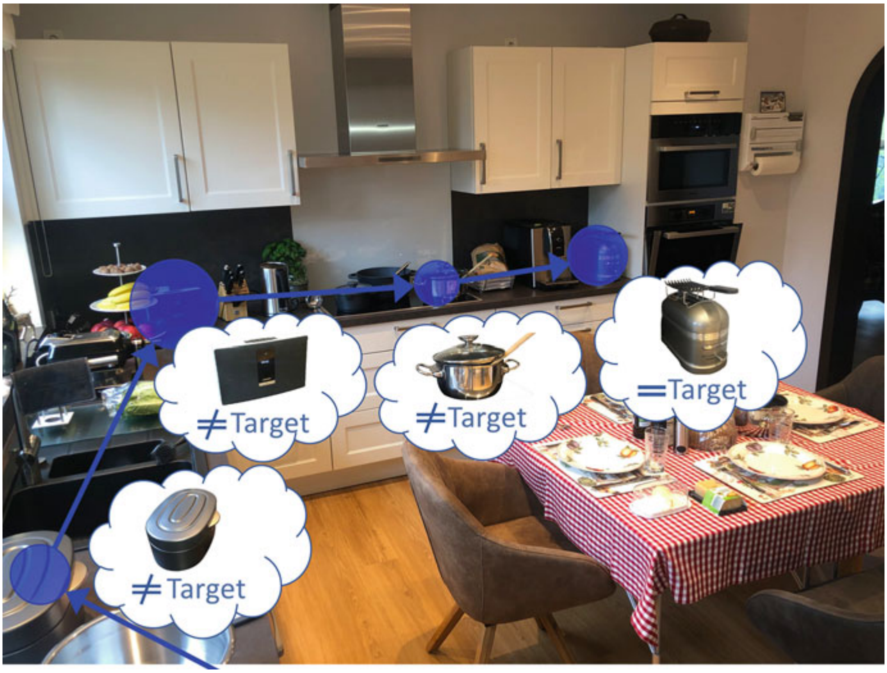


Fig. 4 Illustration of a search for a toaster in the kitchen which not only benefits from contextual guidance of eye-movements but also from faster disengagement of distractors as well as enhanced recognition of the target. Blue arrows and circles illustrate an exemplary scan path and fixations, respectively. Thought bubbles indicate successful object identification and matching against the search template

5.1 Relative Contributions of Context Ingredients

While several context ingredients that modulate search and/or object processing have been identified, relatively little is known about the relative contributions of these ingredients, especially over time. Some work has focused on assessing the relative importance of scene background and object content in visual search, as outlined below. While a brief glance at a scene background without objects was shown to facilitate search, no facilitation was found when local objects were briefly shown instead of the scene background (Võ & Schneider, 2010). Moreover, there is evidence that search performance is higher when scene background versus object content is available throughout the search; yet, both types of information were shown to interact such that scene context provides coarse guidance to relevant regions while object content yields guidance to specific areas (Pereira & Castelhano, 2014). Together, these findings may suggest that global scene properties have a stronger contribution overall, and that they are utilized more readily than local properties—the latter presumably first need to be parsed through a "selective pathway" that binds features into objects (Wolfe et al., 2011b). However, in some other studies, the influence of objects was arguably stronger: Koehler and Eckstein (2017a, 2017b) found scene background to affect search and perceptual decisions less than object content which was divided into a co-occurring object (in close proximity to the target) and multiple object configuration (encompassing all other objects in the scene). While the influence of multiple object configuration was already present in early eye-movements, co-occurring object information was utilized later for a more fine-grained type of guidance. These findings are in line with work showing that object content yields semantic guidance in the absence of scene gist information—but only when the spatial configuration of objects is intact—while scene gist does not enhance the utilization of semantic information when object content is available (Wu et al., 2014a).

One explanation of the mixed results as to the relative importance of scene background could be that scene background is not equally informative of the target's location across studies. Interestingly, Koehler and Eckstein (2017a, 2017b) found that judgments of expected target locations per context ingredient predicted the magnitude of eye-movement guidance for that particular ingredient—the inferior contribution of scene background was related to the finding that scene background was the least informative of the target's location. In other studies, scene background may have been more informative, possibly related to the way that scene background and object content was defined. For instance, larger elements (e.g., a bed) are sometimes considered as objects, given that they can plausibly be moved (Koehler & Eckstein, 2017a, 2017b), whereas they are assumed to belong to the background in other cases (Võ & Schneider, 2010; see also Pereira & Castelhano, 2014). These larger objects may not only provide surfaces for local objects (Castelhano & Krzyś, 2020; Pereira & Castelhano, 2019) but also constitute meaningful subunits in scenes: It has recently been established that anchor objects distinctly contribute to search guidance, yielding stronger facilitation than other semantically related objects (Boettcher et al., 2018) or meaningless cuboids of similar sizes (Helbing et al., 2020). Thus, the relative contributions of context ingredients may strongly depend on the precision of the spatial predictions they yield (see also Eckstein, 2017), which may vary across studies and conceptualizations. In other words, the goal of visual search is to locate something, and we may utilize most those properties that precisely "tell us where to look". Naturally, there are constraints with respect to what information is available when/where in the visual system, and at what cost. That is, a property may be very informative but not yet selected and processed to the extent that it can guide search (Wolfe et al., 2011b).

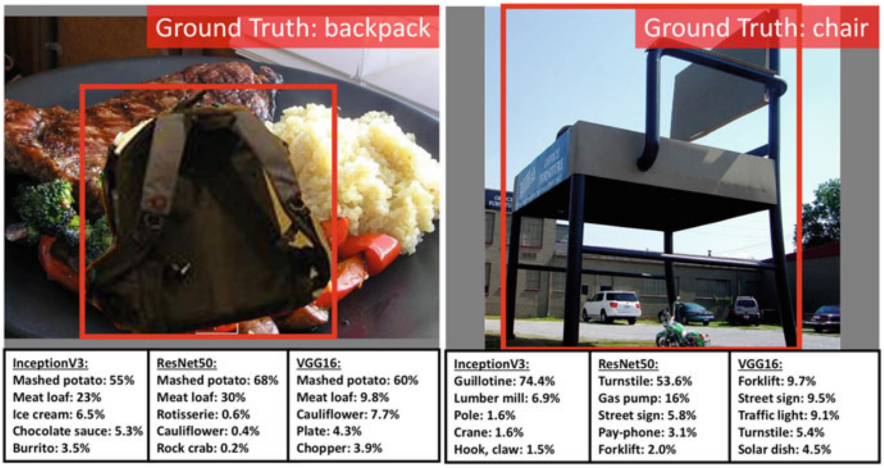
While there is no doubt that spatial predictions are also utilized in object recognition, they are naturally not a prerequisite for contextual modulation of object processing. For instance, facilitation of object recognition is seen even when the target's location is entirely uninformative with respect to its identity (see Lauer et al., 2020a). While the object recognition literature provides clear evidence of contextual modulation in the absence of any recognizable objects in the scene (e.g., Brady et al., 2017), the *relative* contribution of such scene-based (vs. object-based) ingredients has not been assessed in the absence of object content. Currently, there is some indication that both scene context (including objects) and object co-occurrence can yield context effects of a comparable magnitude (Davenport, 2007; Lauer et al., 2020b) and timeline (Lauer et al., 2020b).

Taken together, future work could aim at further teasing apart the relative contributions of distinct global scene properties on the one hand and more local information in the form of various types of co-occurring objects on the other while also assessing how informative they are of the target. It will be especially crucial to test how the various types of scene ingredients exert influence over time.

5.2 *Context in Human and Computer Vision*

Many of the recent advances in better capturing human efficiency in object search and perception in the real world have been facilitated by large-scale databases and computational models (e.g., Boettcher et al., 2018; Greene, 2013, 2016; Rosenholtz et al., 2012)—inspired by a wide range of psychophysical, eye tracking, and neurophysiological studies. Interestingly, in recent years, computer vision algorithms have reached (or even surpassed) human performance levels in a number of tasks. Studying how computational solutions accomplish these tasks may be quite useful for understanding the mechanisms of human visual perception even better in the future. That is, computational models inspired by the visual system can, if validated, be used to test hypotheses about human vision in a highly controlled manner (for a review, see Lindsay, 2020). Despite several challenges, researchers have recently begun to compare and relate DNNs (specifically convolutional neuronal networks, CNNs) to human perception across the hierarchy of the visual system. Activity in the ventral stream has been shown to be generally well predicted by CNNs, with outputs from higher artificial layers better predicting activity in higher visual areas (see Lindsay, 2020). A network trained on scene recognition was able to predict activity in occipital place area, providing insights into the processing of navigational affordances (Bonner & Epstein, 2018; c.f. Lindsay, 2020). There is also electrophysiological evidence of shared spatiotemporal scene category information in humans and DNNs (Greene & Hansen, 2018). On the behavioral level, one cannot only compare benchmark classification accuracy (for which DNNs are commonly optimized) but also error patterns that humans and DNNs might share—or not share (Wichmann et al., 2017). For example, why is it that humans unlike DNNs sometimes fail to notice giant targets in scenes even when they are salient and fixated (Eckstein et al., 2017)? These and other assessments may provide further insights into the mechanisms of search and object perception in scene context. To test or to generate new hypotheses, biologically inspired artificial networks can be altered in many ways, for instance by manipulating their architectures, training sets, or training procedures (see Lindsay, 2020). Of course, careful comparisons and interpretations are important; humans and neuronal networks may achieve a very similar task outcome but accomplish it in an entirely different way computationally.

On the other hand, for large parts of the computer vision community, neuronal networks need not be biologically plausible; they are commonly intended to achieve the highest possible performance in a given tasks such as object recognition. An understanding of human efficiency in object perception, however, may help in further optimizing computer vision algorithms. A key difference between state-of-the-art DNNs and the human visual machinery is that DNNs still require massive quantities of labeled training data, while humans can learn new object concepts from a very small number of examples (Morgenstern et al., 2019; Spiegel & Halberda, 2011). Moreover, while DNNs achieve benchmark performance under certain controlled conditions, they sometimes fail under slightly changing conditions (e.g., image degradation or contrast reduction; Geirhos et al., 2018; Wichmann et al., 2017) and may thus lack the robustness and flexibility of human vision. Intriguingly, misclassification can even occur when the target is altered in a way that is unnoticeable for the human eye (see “adversarial examples”, e.g., Goodfellow et al., 2014). In many algorithms for object recognition, context information is utilized only indirectly (Zhang et al., 2020): For instance, DNNs trained on object recognition in natural scenes typically represent some contextual features implicitly, which becomes apparent when they are fooled by a target-incongruent scene context (see Fig. 5). Recently, Zhang et al. (2020) introduced a biologically-inspired model that builds on feature extraction of a state-of-the-art DNN for object recognition (VGG16), yet incorporates scene context more explicitly; target and context features are processed in parallel using a dual stream architecture, with an attention mechanism selecting informative parts of the context. Performance of the



InceptionV3:	ResNet50:	VGG16:
Mashed potato: 55%	Mashed potato: 68%	Mashed potato: 60%
Meat loaf: 23%	Meat loaf: 30%	Meat loaf: 9.8%
Ice cream: 6.5%	Rotisserie: 0.6%	Cauliflower: 7.7%
Chocolate sauce: 5.3%	Cauliflower: 0.4%	Plate: 4.3%
Burrito: 3.5%	Rock crab: 0.2%	Chopper: 3.9%

InceptionV3:	ResNet50:	VGG16:
Guillotine: 74.4%	Turnstile: 53.6%	Forklift: 9.7%
Lumber mill: 6.9%	Gas pump: 16%	Street sign: 9.5%
Pole: 1.6%	Street sign: 5.8%	Traffic light: 9.1%
Crane: 1.6%	Pay-phone: 3.1%	Turnstile: 5.4%
Hook, claw: 1.5%	Forklift: 2.0%	Solar dish: 4.5%

Fig. 5 DNNs for object recognition do not yield an accurate label when the scene context is inconsistent with the target, as seen in the top-five labels and corresponding confidence levels for three state-of-the art models. This indicates that DNNs represent some contextual features even though they are usually not specifically designed to do so. Figure reproduced from Zhang et al. (2020) with permission

Context aware Two-stream Attention network (CATNet) was highly correlated with human performance across various object recognition experiments (manipulating the quantity, quality and dynamics of context) while also outperforming other models overall. However, it should be noted that all tested models, including CATNet, performed considerably worse than humans when the targets were small. That is, variable object size in natural images still remains a challenge in computer vision (Zhang et al., 2020). Another recent dual pathway model, *GistNet,* was designed such that it would utilize coarse global features of the context, inspired by human scene gist perception (Wu et al., 2018). GistNet was shown to outperform VGG16, even when the scene context was significantly blurred which strongly reduced recognizable objects. The authors also visualized what features the two streams actually utilized and concluded that the foveal pathway employs "local edges and lines", while the global pathway finds "more holistic scene information corresponding to gist-like features" (p. 5).

Taken together, similar to exploring the influence of various context ingredients in human perception, these same ingredients could also be put to a test in computer vision applications furthering the reciprocal benefit when combining methods and theory of both research areas.

6 Conclusion

The way we search for, identify, and interact with objects in the real world is substantially shaped by the scene context in which they occur. In this chapter, we outline recent endeavors to determine what context information (or "ingredients") are actually utilized by the visual system for efficient object localization and identification. We argue that, in both domains, a rapidly acquired coarse global representation of the scene, which can be inferred from spatial layout information, can already lead to contextual modulation. Moreover, at least indoor scenes tend to be organized hierarchically with various levels of context exerting strong influence on both object search and perception. While we have begun to understand which ingredients of a scene matter, there is still much work to be done to more precisely assess the relative contributions of various context ingredients, especially as they unfold over space and time.

Acknowledgments This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project number 222641018—SFB/TRR 135, sub-project C7 to MLV.

References

Auckland, M. E., Cave, K. R., & Donnelly, N. (2007). Nontarget objects can influence perceptual processes during object recognition. *Psychonomic Bulletin & Review, 14*(2), 332–337. https://doi.org/10.3758/BF03194073

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences, 16*(8), 437–443. https://doi.org/10.1016/j.tics.2012.06.010

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

Barrett, S. E., & Rugg, M. D. (1990). Event-related potentials and the semantic matching of pictures. *Brain and Cognition, 14*(2), 201–212. https://doi.org/10.1016/0278-2626(90)90029-N

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

Boettcher, S. E. P., Draschkow, D., Dienhart, E., & Võ, M. L.-H. (2018). Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search. *Journal of Vision, 18*(13), 1–13. 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

Bonner, M. F., & Epstein, R. A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLoS Computational Biology, 14*(4), 1–31. https://doi.org/10.1371/journal.pcbi.1006111

Borji, A. (2019). Saliency prediction in the deep learning era: Successes and limitations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–44. https://doi.org/10.1109/tpami.2019.2935715

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185–207. https://doi.org/10.1109/TPAMI.2012.89

Borji, A., Sihite, D. N., & Itti, L. (2013). What stands out in a scene? A study of human explicit saliency judgment. *Vision Research, 91*, 62–77. https://doi.org/10.1016/j.visres.2013.07.016

Boyce, S. J., & Pollatsek, A. (1992). Identification of objects in scenes: The role of scene background in object naming. *Journal of Experimental Psychology, Learning, Memory, and Cognition, 18*(3), 531–543. https://doi.org/10.1037/0278-7393.18.3.531

Boyce, S. J., Pollatsek, A., & Rayner, K. (1989). Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 556–566. https://doi.org/10.1037/0096-1523.15.3.556

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

Brandman, T., Avancini, C., Leticevscaia, O., & Peelen, M. V. (2019). Auditory and semantic cues facilitate decoding of visual object category in MEG. *Cerebral Cortex*, 1–28. https://doi.org/10.1093/cercor/bhz110

Brandman, T., & Peelen, M. V. (2017). Interaction between scene and object processing revealed by human fMRI and MEG decoding. *The Journal of Neuroscience, 37*(32), 7700–7710. https://doi.org/10.1523/jneurosci.0582-17.2017

Brandman, T., & Peelen, M. V. (2019). Signposts in the fog: Objects facilitate scene representations in left scene-selective cortex. *Journal of Cognitive Neuroscience, 31*(3), 390–400. https://doi.org/10.1162/jocn_a_01258

Castelhano, M. S., & Heaven, C. (2011). Scene context influences without scene gist: Eye movements guided by spatial associations in visual search. *Psychonomic Bulletin and 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., & Krzyś, K. (2020). Rethinking space: A review of perception, attention, and memory in scene processing. *Annual Review of Vision Science, 6*, 563–586. https://doi.org/10.1146/annurev-vision-121219-081745

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience, 17*(3), 455–462. https://doi.org/10.1038/nn.3635

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

Coco, M. I., Nuthmann, A., & Dimigen, O. (2019). 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. L.-H. (2017). Stuck on semantics: Processing of irrelevant object-scene inconsistencies modulates ongoing gaze behavior. *Attention, Perception, & Psychophysics, 79*(1), 154–168. https://doi.org/10.3758/s13414-016-1203-7

Davenport, J. L. (2007). Consistency effects between objects in scenes. *Memory & Cognition, 35*(3), 393–401. https://doi.org/10.3758/BF03193280

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

De Graef, P., Christiaens, D., & Ydewalle, G. (1990). Perceptual effect of scene context on object identification. *Psychological Research*, 317–329. https://doi.org/10.1007/BF00868064

De Groot, F., Huettig, F., & Olivers, C. N. L. (2016). When meaning matters: The temporal dynamics of semantic influences on visual attention. *Journal of Experimental Psychology: Human Perception and Performance, 42*(2), 180–196. https://doi.org/10.1037/xhp0000102

Draschkow, D., Heikel, E., Võ, M. L.-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

Eckstein, M. P. (2017). Probabilistic computations for attention, eye movements, and search. *Annual Review of Vision Science, 3*, 319–342. https://doi.org/10.1146/annurev-vision-102016-061220

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

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision, 8*(14), 1–26. https://doi.org/10.1167/8.14.18

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision, 8*(3), 1–15. https://doi.org/10.1167/8.3.3

Epstein, R. A., & Baker, C. I. (2019). Scene perception in the human brain. *Annual Review of Vision Science, 5*(1), 373–397. https://doi.org/10.1146/annurev-vision-091718-014809

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

Furtak, M., Doradzińska, Ł., Ptashynska, A., Mudrik, L., Nowicka, A., & Bola, M. (2020). Automatic attention capture by threatening, but not by semantically incongruent natural scene images. *Cerebral Cortex, 30*(7), 4158–4168. https://doi.org/10.1093/cercor/bhaa040

Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Cognitive Brain Research, 16*, 123–144. https://doi.org/10.1016/s0926-6410(02)00244-6

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ICLR*, 1–20. https://zhuanlan.zhihu.com/p/81257789%0Ahttps://github.com/rgeirhos/texture-vs-shape%0Ahttps://github.com/rgeirhos/Stylized-ImageNet 3.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. In *3rd international conference on learning representations, ICLR 2015 – conference track proceedings* (pp. 1–11) http://arxiv.org/abs/1412.6572

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. (2016). Estimations of object frequency are frequently overestimated. *Cognition, 149*, 6–10. https://doi.org/10.1016/j.cognition.2015.12.011

Greene, M. R., & Hansen, B. C. (2018). Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLoS Computational Biology, 14*(7), 1–17. https://doi.org/10.1371/journal.pcbi.1006327

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

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

Gronau, N. (2020). Vision at a glance: The role of attention in processing object-to-object categorical relations. *Attention, Perception, and Psychophysics, 82*(2), 671–688. https://doi.org/10.3758/s13414-019-01940-z

Gronau, N., & Shachar, M. (2014). Contextual integration of visual objects necessitates attention. *Attention, Perception, and Psychophysics, 76*(3), 695–714. https://doi.org/10.3758/s13414-013-0617-8

Helbing, J., Draschkow, D., & Võ, M. L.-H. (2020). Semantic and syntactic anchor object information interact to make visual search in immersive scenes efficient. *Journal of Vision, 20*(11), 573. https://doi.org/10.1167/jov.20.11.573

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science, 16*(4), 219–222. https://doi.org/10.1111/j.1467-8721.2007.00507.x

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. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision, 18*(6), 1–18. https://doi.org/10.1167/18.6.10

Henderson, J. M., Hayes, T. R., Peacock, C. E., & Rehrig, G. (2019). Meaning and attentional guidance in scenes: A review of the meaning map approach. *Vision, 3*(2), 1–10. https://doi.org/10.3390/vision3020019

Henderson, J. M., Hayes, T. R., Rehrig, G., & Ferreira, F. (2018). Meaning guides attention during real-world scene description. *Scientific Reports, 8*(1), 1–9. https://doi.org/10.1038/s41598-018-31894-5

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

Henderson, J. M., Pollatsek, A., & Rayner, K. (1987). Effects of foveal priming and Extrafoveal preview on object identification. *Journal of Experimental Psychology: Human Perception and Performance, 13*(3), 449–463. https://doi.org/10.1037/0096-1523.13.3.449

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

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General, 127*(4), 398–415. https://doi.org/10.1037/0096-3445.127.4.398

Hollingworth, A., & Henderson, J. M. (1999). Object identification is isolated from scene semantic constraint: Evidence from object type and token discrimination. *Acta Psychologica, 102*(2–3), 319–343. https://doi.org/10.1016/S0001-6918(98)00053-5

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. (2007). Visual salience. *Scholarpedia, 2*(9), 3327. https://doi.org/10.4249/scholarpedia.3327

Itti, L., & Baldi, P. (2005). Bayesian surprise attracts human attention. *Advances in Neural Information Processing Systems*, 547–554. https://doi.org/10.1016/j.visres.2008.09.007

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

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*(3), 194–203. https://doi.org/10.1038/35058500

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research, 47*(26), 3286–3297. https://doi.org/10.1016/j.visres.2007.09.013

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

Koehler, K., & Eckstein, M. P. (2017a). Beyond scene gist: Objects guide search more than scene background. *Journal of Experimental Psychology: Human Perception and Performance, 43*(6), 1177–1193. https://doi.org/10.1037/xhp0000363

Koehler, K., & Eckstein, M. P. (2017b). Temporal and peripheral extraction of contextual cues from scenes during visual search. *Journal of Vision, 17*(2), 1–32. https://doi.org/10.1167/17.2.16

Kotseruba, I., Wloka, C., Rasouli, A., & Tsotsos, J. K. (2020). Do saliency models detect odd-one-out targets? *New Datasets and Evaluations*, 1–14. http://arxiv.org/abs/2005.06583

Kovalenko, L. Y., Chaumon, M., & Busch, N. A. (2012). A pool of pairs of related objects (POPORO) for investigating visual semantic integration: Behavioral and electrophysiological validation. *Brain Topography, 25*(3), 272–284. https://doi.org/10.1007/s10548-011-0216-8

Krasovskaya, S., & Macinnes, W. J. (2019). Salience models: A computational cognitive neuroscience review. *Vision, 3*(4). https://doi.org/10.3390/vision3040056

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. *ArXiv*, 1–16. http://arxiv.org/abs/1610.01563

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*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*, 203–205. https://doi.org/10.1126/science.7350657

Kutas, M., & Hillyard, S. A. (1983). Event-related brain potentials to grammatical errors and semantic anomalies. *Memory & Cognition, 11*(5), 539–550. https://doi.org/10.3758/BF03196991

Lauer, T., Boettcher, S. E. P., Kollenda, D., Draschkow, D., & Võ, M. L.-H. (2020b). Manipulating semantic consistency between two objects and a scene: An ERP paradigm. *Journal of Vision, 20*(11), 1078. https://doi.org/10.1167/jov.20.11.1078

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), 1–12. https://doi.org/10.1038/s41598-018-32991-1

Lauer, T., Willenbockel, V., Maffongelli, L., & Võ, M. L.-H. (2020a). The influence of scene and object orientation on the scene consistency effect. *Behavioural Brain Research, 394*, 1–13. https://doi.org/10.1016/j.bbr.2020.112812

Leroy, A., Faure, S., & Spotorno, S. (2020). Reciprocal semantic predictions drive categorization of scene contexts and objects even when they are separate. *Scientific Reports, 10*(1), 1–13. https://doi.org/10.1038/s41598-020-65158-y

Li, B., Gao, C., & Wang, J. (2019). Electrophysiological correlates of masked repetition and conceptual priming for visual objects. *Brain and Behavior, 9*(10), 1–8. https://doi.org/10.1002/brb3.1415

Lindsay, G. W. (2020). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, 1–15. https://doi.org/10.1162/jocn_a_01544

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance, 4*(4), 565–572. https://doi.org/10.1037/0096-1523.4.4.565

Loschky, L. C., Szaffarczyk, S., Beugnet, C., Young, M. E., & Boucart, M. (2019). The contributions of central and peripheral vision to scenegist recognition with a 180° visual field. *Journal of Vision, 19*(5), 1–21. https://doi.org/10.1167/19.5.15

MacEvoy, S. P., & Epstein, R. A. (2011). Constructing scenes from objects in human occipitotemporal cortex. *Nature Neuroscience, 14*(10), 1323–1329. https://doi.org/10.1038/nn.2903

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), 1–16. 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

Masciocchi, C. M., Mihalas, S., Parkhurst, D., & Niebur, E. (2009). Everyone knows what is interesting: Salient locations which should be fixated. *Journal of Vision, 9*(11), 1–22. https://doi.org/10.1167/9.11.1

McPherson, W. B., & Holcomb, P. J. (1999). An electrophysiological investigation of semantic priming with pictures of real objects. *Psychophysiology, 36*(1), 53–65. https://doi.org/10.1017/S0048577299971196

Morgenstern, Y., Schmidt, F., & Fleming, R. W. (2019). One-shot categorization of novel object classes in humans. *Vision Research, 165*, 98–108. https://doi.org/10.1016/j.visres.2019.09.005

Mudrik, L., Lamy, D., & Deouell, L. Y. (2010). ERP evidence for context congruity effects during simultaneous object-scene processing. *Neuropsychologia, 48*, 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

Munneke, J., Brentari, V., & Peelen, M. V. (2013). The influence of scene context on object recognition is independent of attentional focus. *Frontiers in Psychology, 4*, 1–10. https://doi.org/10.3389/fpsyg.2013.00552

Neider, M. B., & Zelinsky, G. J. (2006). 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., De Groot, F., Huettig, F., & Olivers, C. N. L. (2019). Extrafoveal attentional capture by object semantics. *PLoS One, 14*(5), 1–19. https://doi.org/10.1371/journal.pone.0217051

Nuthmann, A., & Einhäuser, W. (2015). A new approach to modeling the influence of image features on fixation selection in scenes. *Annals of the New York Academy of Sciences, 1339*(1), 82–96. https://doi.org/10.1111/nyas.12705

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision, 10*(8), 1–19. https://doi.org/10.1167/10.8.20

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*, 145–175. https://doi.org/10.1023/A:1011139631724

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research, 155*, 23–36. https://doi.org/10.1016/S0079-6123(06)55002-2

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

Onat, S., Açik, A., Schumann, F., & König, P. (2014). The contributions of image content and behavioral relevancy to overt attention. *PLoS One, 9*(4). https://doi.org/10.1371/journal.pone.0093254

Pajak, M., & Nuthmann, A. (2013). Object-based saccadic selection during scene perception: Evidence from viewing position effects. *Journal of Vision, 13*(5), 1–21. https://doi.org/10.1167/13.5.2

Palmer, T. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition, 3*(5), 519–526. https://doi.org/10.3758/BF03197524

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*(1), 107–123. https://doi.org/10.1016/S0042-6989(01)00250-4

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019a). Meaning guides attention during scene viewing, even when it is irrelevant. *Attention, Perception, and Psychophysics, 81*(1), 20–34. https://doi.org/10.3758/s13414-018-1607-7

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019b). The role of meaning in attentional guidance during free viewing of real-world scenes. *Acta Psychologica, 198*(July). https://doi.org/10.1016/j.actpsy.2019.102889

Pedziwiatr, M. A., Wallis, T. S. A., Kümmerer, M., & Teufel, C. (2019). Meaning maps and deep neural networks are insensitive to meaning when predicting human fixations. *Journal of Vision, 19*(10), 253c. https://doi.org/10.1101/840256.

Pereira, E. J., & Castelhano, M. S. (2014). Peripheral guidance in scenes: The interaction of scene context and object content. *Journal of Experimental Psychology: Human Perception and Performance, 40*(5), 2056–2072. https://doi.org/10.1037/a0037524

Pereira, E. J., & Castelhano, M. S. (2019). Attentional capture is contingent on scene region: Using surface guidance framework to explore attentional mechanisms during search. *Psychonomic Bulletin and Review, 26*(4), 1273–1281. https://doi.org/10.3758/s13423-019-01610-z

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40*, 49–71. https://doi.org/10.1023/A:1026553619983

Quek, G. L., & Peelen, M. V. (2020). Contextual and spatial associations between objects interactively modulate visual processing. *Cerebral Cortex*, 1–14. https://doi.org/10.1093/cercor/bhaa197

Rehrig, G., Hayes, T. R., Henderson, J. M., & Ferreira, F. (2020). When scenes speak louder than words: Verbal encoding does not mediate the relationship between scene meaning and visual attention. *Memory and Cognition, 48*(7), 1181–1195. https://doi.org/10.3758/s13421-020-01050-4

Roberts, K. L., & Humphreys, G. W. (2011). Action relations facilitate the identification of briefly-presented objects. *Attention, Perception, and Psychophysics, 73*(2), 597–612. https://doi.org/10.3758/s13414-010-0043-0

Rosenholtz, R., Huang, J., & Ehinger, K. (2012). Rethinking the role of top-down attention in vision: Effects attributable to a lossy representation in peripheral vision. *Frontiers in Psychology, 3*(FEB), 1–15. https://doi.org/10.3389/fpsyg.2012.00013

Rousselet, G. A., Joubert, O. R., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition, 12*(6), 852–877. https://doi.org/10.1080/13506280444000553

Roux-Sibilon, A., Trouilloud, A., Kauffmann, L., Guyader, N., Mermillod, M., & Peyrin, C. (2019). Influence of peripheral vision on object categorization in central vision. *Journal of Vision, 19*(14), 1–16. https://doi.org/10.1167/19.14.7

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision, 77*, 157–173. https://doi.org/10.1007/s11263-007-0090-8

Sastyin, G., Niimi, R., & Yokosawa, K. (2015). Does object view influence the scene consistency effect? *Attention, Perception, & Psychophysics, 77*, 856–866. https://doi.org/10.3758/s13414-014-0817-x

Schomaker, J., Walper, D., Wittmann, B. C., & Einhäuser, W. (2017). Attention in natural scenes: Affective-motivational factors guide gaze independently of visual salience. *Vision Research, 133*, 161–175. https://doi.org/10.1016/j.visres.2017.02.003

Schütt, H. H., Rothkegel, L. O. M., Trukenbrod, H. A., Engbert, R., & Wichmann, F. A. (2019). Disentangling bottom-up versus top-down and low-level versus high-level influences on eye movements over time. *Journal of Vision, 19*(3), 1–23. https://doi.org/10.1167/19.3.1

Spain, M., & Perona, P. (2011). Measuring and predicting object importance. *International Journal of Computer Vision, 91*(1), 59–76. https://doi.org/10.1007/s11263-010-0376-0

Spiegel, C., & Halberda, J. (2011). Rapid fast-mapping abilities in 2-year-olds. *Journal of Experimental Child Psychology, 109*(1), 132–140. https://doi.org/10.1016/j.jecp.2010.10.013

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

t'Hart, B. M., Schmidt, H. C. E. F., Roth, C., & Einhäuser, W. (2013). Fixations on objects in natural scenes: Dissociating importance from salience. *Frontiers in Psychology, 4*, 1–9. https://doi.org/10.3389/fpsyg.2013.00455

Teufel, C., & Fletcher, P. C. (2020). Forms of prediction in the nervous system. *Nature reviews neuroscience, 21*(4), 231–242. https://doi.org/10.1038/s41583-020-0275-5

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., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology, 59*(11), 1931–1949. https://doi.org/10.1080/17470210500416342

Underwood, G., Humphreys, L., & Cross, E. (2007). Congruency, saliency and gist in the inspection of objects in natural scenes. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 563–579). Elsevier. https://doi.org/10.1016/B978-008044980-7/50028-8

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

Võ, M. L.-H. (2021). The meaning and structure of scenes. *Vision Research, 181*, 10–20. https://doi.org/10.1016/j.visres.2020.11.003

Võ, M. L.-H., Boettcher, S. E., & 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., & Henderson, J. M. (2009a). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision, 9*(3), 1–15. https://doi.org/10.1167/9.3.24

Võ, M. L.-H., & Henderson, J. M. (2009b). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision, 9*(3), 1–15. 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), 1–13. https://doi.org/10.1167/10.3.14

Võ, M. L.-H., & Henderson, J. M. (2011). Object-scene inconsistencies do not capture gaze: Evidence from the flash-preview moving-window paradigm. *Attention, Perception, and Psychophysics, 73*(6), 1742–1753. https://doi.org/10.3758/s13414-011-0150-6

Võ, M. L.-H., & Schneider, W. X. (2010). A glimpse is not a glimpse: Differential processing of flashed scene previews leads to differential target search benefits. *Visual Cognition, 18*(2), 171–200. https://doi.org/10.1080/13506280802547901

Võ, M. L.-H., & Wolfe, J. M. (2013a). Differential electrophysiological signatures of semantic and syntactic scene processing. *Psychological Science, 24*, 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. L.-H., & Wolfe, J. M. (2015). The role of memory for visual search in scenes. *Annals of the New York Academy of Sciences, 1339*(1), 72–81. https://doi.org/10.1111/nyas.12667

Wichmann, F. A., Janssen, D. H. J., Geirhos, R., Aguilar, G., Schütt, H. H., Maertens, M., & Bethge, M. (2017). Methods and measurements to compare men against machines. *Electronic Imaging, 2017*(14), 36–45. https://doi.org/10.2352/ISSN.2470-1173.2017.14.HVEI-113

Wolfe, J. M. (2020). Visual search: How do we find what we are looking for? *Annual Review of Vision Science, 6*, 539–562. https://doi.org/10.1146/annurev-vision-091718-015048

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, and Psychophysics, 73*(6), 1650–1671. https://doi.org/10.3758/s13414-011-0153-3

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour, 1*(3), 1–8. https://doi.org/10.1038/s41562-017-0058

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., Wang, H. C., & Pomplun, M. (2014a). The roles of scene gist and spatial dependency among objects in the semantic guidance of attention in real-world scenes. *Vision Research, 105*, 10–20. https://doi.org/10.1016/j.visres.2014.08.019

Wu, C. C., Wick, F. A., & Pomplun, M. (2014b). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology, 5*, 1–13. https://doi.org/10.3389/fpsyg.2014.00054

Wu, K., Wu, E., & Kreiman, G. (2018). Learning scene gist with convolutional neural networks to improve object recognition. *ArXiv*, 1–6. http://arxiv.org/abs/1803.01967

Zhang, M., Tseng, C., & Kreiman, G. (2020). Putting visual object recognition in context. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020*, 12982–12991. https://doi.org/10.1109/CVPR42600.2020.01300

Zucker, L., & Mudrik, L. (2019). Understanding associative vs. abstract pictorial relations: An ERP study. *Neuropsychologia, 133*, 1–16. https://doi.org/10.1016/j.neuropsychologia.2019.107127