CHAPTER EIGHT

Changing perspectives on goal-directed attention control: The past, present, and future of modeling fixations during visual search

Gregory J. Zelinsky[a,b,*], Yupei Chen[a], Seoyoung Ahn[a], and Hossein Adeli[a]

[a]Department of Psychology, Stony Brook University, New York, NY, United States
[b]Department of Computer Science, Stony Brook University, New York, NY, United States
*Corresponding author: e-mail address: gregory.zelinsky@stonybrook.edu

Contents

Psychology of Learning and Motivation, Volume 73
ISSN 0079-7421
https://doi.org/10.1016/bs.plm.2020.08.001

Abstract

People make eye movements while interacting with objects, and these behaviors are rich with information about how visual goals are represented in the brain and used to prioritize sequential motor behavior. Here we adopt a real-world perspective and define goal-directed attention control as the guidance (or biasing) of gaze to target-object goals that have uncertain visual appearance. Specifically, we review models of goal-directed attention control that have attempted to predict the behavioral fixations made in the search for target-category goals in images. We will show how modeling perspectives on this question changed over the decades. Using the year 2020 as a reference, we will critically review the recent past of the categorical search modeling literature (∼2000–2010), the literature defining our present (∼2010–2020), and speculate about the future of search models and the directions that the literature may turn in the next decade (∼2020–2030).

1. Defining the problem

In this section we aim to define the scope of our question, provide some justification for why we defined it as we did, and argue for why asking this question is important. Critical to a review is a clear definition of the specific question under consideration, and the situation of this question in related literatures that are specifically not considered so as to create a context. Our focus is on the goal-directed attention literature that attempts to predict the fixations made by people searching for common object goals in contexts that approximate the real world. In this section we elaborate on two of the three key components in this problem definition. First, that visual search is an ideal paradigm for studying goal-directed attention control. This is particularly true for categorical visual search, which, contrary to being a contrived laboratory task, is a cognitively meaningful goal-directed behavior that we engage in probably hundreds of times each day. Second, that these models must be able to predict changes in gaze fixations, which we consider to be the most basic observable behavioral measure of visuo-spatial attention control. We also consider fixation prediction to be a desirable level of granularity with which to study goal-directed behavior, and believe it to be an achievable modeling goal given the methods currently available to modelers. The third component to our definition is that this prediction of search fixations must be in a reasonably realistic context. We elaborate on our definition of a realistic context in a section that we devote to the past, but we believe part of this definition must include some non-trivial uncertainty about the target's appearance.

1.1 Visual search: The simplest of goal-directed behaviors

Goal-directed attention control underlies all the things that we *try* to do. Whether it is making dinner or navigating across town, the vast majority of tasks that we perform require executing a coordinated sequence of goals that produce coordinated sequences of motor behaviors. Goal-directed control also implies that different goals require different behavior. The arm control needed to drive a car would be of little use if the goal was to swim, and vice versa. Researchers have studied goal-directed attention control for decades, with classic demonstrations dating back to Yarbus (1967). Early theoretical work also focused on the role of attention in controlling and coordinating goal-directed behavior (e.g., Broadbent, 1957; Norman & Shallice, 1986), with task performance believed to reflect the degree and efficiency that this control can be enacted (Allport, 1980). More recently, the study of goal-directed behavior has extended to highly naturalistic contexts and tasks, such as driving (Land, 1992), making a sandwich or a cup of tea (Land & Hayhoe, 2001), and taking a walk (Foulsham, Walker, & Kingstone, 2011). Among the many interesting findings from these studies is that even seemingly simple tasks are nevertheless accompanied by highly coordinated and complex movements of gaze fixation (Land & Tatler, 2009), and this complexity challenges a computational understanding of how they all relate to a fluidly changing goal state. For the purpose of developing a tractable model the aim is in some sense the opposite, which is to identify the goal-directed behaviors of *least* complexity. There are still several candidates from which to choose, depending on one's definition of a goal. Even simple demonstrations of object-based attention (Scholl, 2001) can be considered goal-directed control if the goal is to perceive an object and the control is the top-down object bias. Our work has focused on visual search, which explains in part our problem definition. We chose search because it is arguably the simplest, and therefore best, goal-directed behavior to model. In a visual search task there is a target goal and the task is to find it. This simplicity, combined with its clear expression in eye movement behavior, makes visual search a valuable paradigm for studying goal-directed attention control. If you show someone a kitchen scene and ask them to find the microwave (Fig. 1A), you see a very different distribution of fixations than if you ask them to find the clock (Fig. 1B). Critically, these different behaviors are observed despite the visual input being identical. This demonstration, and others like it, provide proof positive of the top-down, goal-directed nature of the search task. Search researchers refer to this as *target guidance*

Fig. 1 The goal-directed control of gaze during visual search. (A) Fixation-density map for a clock search. (B) Fixation-density map for a microwave search. (C) Saliency map from the DeepGaze II model.

(Wolfe, 1994), with the overt variety of guidance considered here being the biasing of eye movements to the target goal by top-down attention control (Zelinsky, 2008). Visual search is also a natural task, not a contrived laboratory paradigm, and it is fertile ground for the study of interacting top-down and bottom-up processes and the information routing that leads up to object recognition. Unlike an anti-saccade task (Munoz & Everling, 2004), or even the attention control needed to overcome Stroop interference or other automated responses (Wright, 2017), visual search is a goal-directed behavior that has a connection to the real world that these other tasks do not. We are able to search for visually-complex objects, and even object categories. It is difficult to imagine how other, comparably simple paradigms could be scaled up to a real-world context, or how they could be modeled. Finally, visual search is very theory rich, and this strong modeling literature has taken up momentum over the last couple of decades, as we hope to show in this review.

Visual search is not free viewing, and models of one are not models of the other. Indeed, if visual search is the quintessential goal-directed task, its antithesis is free viewing, the quintessential "taskless" viewing behavior (see Borji (2019) for a recent review of work using a free-viewing paradigm). There has been impressive growth in the modeling literature aimed at predicting attention control in the context of free viewing, with the most salient among these being the seminal work by Itti and colleagues (Itti & Koch, 2001; Itti, Koch, & Niebur, 1998). They computed a feature-contrast signal that they termed *salience*, and used saliency models to predict the role of the bottom-up visual input in controlling the direction of spatial attention. This focus on bottom-up control led naturally to use of the free-viewing paradigm, which is a good marriage of methodology and modeling objective. However, despite free-viewing fixation prediction being an

active literature where there are even competitions and leaderboards for the best models (https://saliency.tuebingen.ai/), saliency models are fundamentally not models of goal-directed attention. It therefore follows that saliency models are poor predictors of search fixation behavior (Chen & Zelinsky, 2006; Henderson, Brockmole, Castelhano, & Mack, 2007; Koehler, Guo, Zhang, & Eckstein, 2014). Fig. 1C drives home this fact by showing that even one of the best saliency models (Kümmerer, Wallis, & Bethge, 2016) cannot predict the goal-specific biasing of gaze that is at the heart of overt visual search. We therefore will not be considering this literature in this review, other than in the context of the datasets and metrics that were developed in the saliency modeling literature that have applicability to models of search. To clarify this distinction from bottom-up saliency models, we will adopt throughout this review the terminology from Zelinsky and Bisley (2015), which referred to an image prioritization based on bottom-up factors as a *saliency map*, and an image prioritization based on a top-down target goal as a *target map*. Just as visual search is very different from free viewing, target maps are not saliency maps, and vice versa. We reserve the term *priority map* for models that combine this top-down and bottom-up information, or when referring to the more general concept of prioritization where the source of the bias is immaterial.

1.2 Predicting fixations: The higher bar

Researchers who choose to adopt oculomotor dependent measures in their studies of search are often reminded that eye movements and shifts of spatial attention are not the same thing. We get it. Yet, in this review we specifically exclude models that predict only manual button press responses (and more often just group means; e.g., (Bundesen, 1990; Müller, Heller, & Ziegler, 1995; Wolfe, 1994) and models that predict only modulatory effects of attention and not the expression of this modulation in behavior (e.g., Deco & Rolls, 2004; Hamker, 2004, 2005; Reynolds, Pasternak, & Desimone, 2000). We do this because these studies are outside the scope of our problem definition, but this begs the question of why we defined the problem in exactly this way. There are several justifications for this reversal of focus on the search fixations (Zelinsky, 2008), but here we will give three that have stood the test of time.

First, for researchers interested in the mechanism of attention, the debate over the exact relationship between eye movements and shifts of spatial attention is important and lives on (Hunt, Reuther, Hilchey, & Klein, 2019).

However, for the majority of cognitive scientists this debate is over, as indicated by the explosive growth in the use of oculomotor measures of attention over the last decade (Findlay, 2004). This widespread embrace of oculomotor measures reflects a realization that eye movements are highly correlated with shifts of attention, even in the laboratory (Williams, Reingold, Moscovitch, & Behrmann, 1997; Zelinsky & Sheinberg, 1997), and when considered in unconstrained real-world contexts differences between the two become negligible (Findlay, 2005; Findlay & Gilchrist, 2003; Itti, Rees, & Tsotsos, 2005). For modelers whose focus is on the prediction of attention control, what this correlation means is that the priority map used to guide gaze during search is likely a good estimate of the priority map used by purely covert attention to control the routing of inputs and information flow through the visual system. Second, eye movements are explicit in a way that attention shifts are not. Each eye movement is a behavior, an observable decision about where to reposition the high-resolution fovea, the visual system's one indisputable limited resource. Collectively, these saccadic eye movements constitute our most frequent behavior, with 4-5 occurring every second. If this was not already sufficient motivation for a behavioral scientist, eye movements, and the brief periods of relative eye immobility between each known as gaze fixations, have long been thought of as a window into high-level cognitive processes, such as reading (Clifton et al., 2016) and scene understanding (Henderson, 2003). This window arguably sheds the most light on the understanding of goal-directed attention, where each of these little behaviors is made in the service of a top-down goal. Returning to the scenario suggested in Fig. 1, the eye movements that you make while walking into your kitchen will depend on whether your goal is to learn the time or to re-heat a tea. Indeed, tasks can be inferred if only eye-movement information is available with the visual input (Henderson, Shinkareva, Wang, Luke, & Olejarczyk, 2013; Zelinsky, Peng, & Samaras, 2013). The relatively high frequency of eye movements means that sequences of fixations can be observed during the execution of a goal-directed behavior, making possible the characterization of the behavior at a level of spatio-temporal resolution that far exceeds a manual response. In the context of search, instead of obtaining only a single measure of when a search ends, with eye movement paradigms you get to see how the search unfolds over time. Until there is a machine that can measure the movements of an attention-routing window over an image, eye trackers will remain a valuable experimental tool, perhaps the most valuable to behavioral science methodology since the birth of chronometry (Posner, 1978). Third, and following from the previous, the spatial locations of fixations made during

search, and their temporal order, create a highly challenging dataset for model development and testing. Most modeling work on search has aimed at reproducing behavioral trends or neuro-modulatory responses, and not at predicting the locations of search fixations or their sequence. Here, we borrow the bar set by models in the saliency literature by requiring models of goal-directed control to make predictions at this finer-grained level—the locations of individual fixations in an image. For search, this is done by using target maps to predict the sequence of image locations that are fixated en route to a target, much like saliency maps are used to predict fixations during free viewing.

2. Overview

With the problem well defined, we now turn to our review of the existing literature. Adhering to the problem definition that we have justified, our goal was to attempt a comprehensive review of models satisfying that definition. We sincerely wish to apologize to the makers of the models that we undoubtedly missed. We adopt in this review a historical perspective spanning the last two decades, and speculate on the decade in front of us. However, more essential than chronology are the ideas that shaped the research on search modeling, and how these ideas changed over time as the field matured. We aim to use these methodological advances to carve a past from the present with respect to modeling approaches. Whenever possible, we will also use our own work to illustrate examples of models that have outlived their usefulness by relying on outdated methods and assumptions that would limit their capacity for true behavior prediction. The keen-eyed reader will also notice that studies falling into the modeling "present" may seem heavily biased to highlight work from our lab. We realize this appearance, but contend that it is a fair application of the criteria defining the problem. Given these criteria, the number of models in this review is not extensive, so each can be described in some depth. In doing this, it will unfortunately be necessary to introduce some jargon, particularly when describing the methods used by a model or the dataset on which it was trained. Our aim was to provide a context that can help the reader understand broadly what these methods do, but not to de-focus the discussion with too many details. In some sense, the methods and datasets are not the most important things anyway, because they both have typically short lifespans. In all cases, the interested reader can learn more about each method by doing a quick Internet search, and the very interested reader should

consult the original sources for full details. Finally, we will speculate as to some future directions that the modeling literature might want to take, one of which is the importance of developing large-scale datasets of behaviorally-annotated images for training the search models of the future. In this context, we will also describe some very recent work from our lab.

3. The past (~2000–2010)

By what criteria should previous models of visual search, irrespective of their contribution and success, no longer be considered contemporary? Model methods, after all, are constantly improving. These technical innovations can come from many different literatures, and they can happen fast. Deep networks are one example of this, and we will have more to say about this topic later. Models of visual search have not been unaffected by these methodological currents, and this is particularly true for models aimed at predicting search fixations. In some sense the past must be defined by outdated methods and not just a number of years, although the two are correlated. In particular, we consider the following three recent milestones in our carving of past from present in the modeling of search fixations.

3.1 Does your model use visually-complex targets in visually-complex images as search stimuli?

If your answer to this question is "no," your model may be in need of updated methods. Models of search can be designed to work extremely well if they get to use carefully hand-crafted features, or if they can assume highly predictable visual contexts. For example, the Guided Search models (versions GS2 to GS6, Wolfe, 2020; Wolfe, Cain, Ehinger, & Drew, 2015; Wolfe, 1994; Wolfe & Gancarz, 1997; Wolfe, Horowitz, Palmer, Michod, & Van Wert, 2010) were hugely influential in focusing the literature on target guidance, which became a core construct in our understanding of visual search. But this contribution of the GS model is now quite old, and newer versions of the model inherited weak modeling methods from the previous millennium, namely an input that comes discretized into only a handful of feature values. To be clear, we do not see GS's reliance on a narrow feature input as a weakness, but rather that the GS models were a product of their time. In contrast, another model introduced within the GS range of years, the Target Acquisition Model (TAM; Zelinsky, 2008), accepted images as inputs. This means that TAM worked when patterns were defined by a small number of features, but also for all the vagaries of

patterns that can exist in natural images. On a measure of model applicability, here defined as the range of search stimuli to which a model can be applied, TAM would therefore be preferred to GS.

It is perhaps useful to think of this problem as one varying along a dimension of scale. If a problem can be scaled down to a small enough feature dimension, then simple heuristics can be identified to find solutions, such as distinguishing targets from non-targets to program eye movements. The problem arises when this solution is scaled up to include greater variability in visual context. When this happens, the simple heuristics used by the scaled-down model often break, resulting in poor predictions. Model-heavy literatures such as computer vision take scalability very seriously, again because it directly impacts the range of tasks to which a model can be applied. Models in the psychology literature tend to take scalability and generalizability less seriously, and all too often adopt an approach that seems based on the idea of a promissory note. Modelers ask researchers to trust that a model developed at a simple scale will also work when scaled up to more realistic inputs, but these promises are rarely kept. Modelers offer three common retorts to this admittedly harsh criticism. One is that interesting research questions can be framed purely within the scaled-down scope of a model, and this is fair. However, it is also fair to ask why a model that is specific to a simpler scale is interesting or important, given that that scale is very different from the visual system's everyday input. A second argument is that such demonstrations of scalability are beyond the scope of search models, which rightly should be focused on search. It is true that scaling up to images introduces a host of highly non-trivial problems, such as the potential for occlusion, variability in perspective, figure/ground segmentation, etc., each of which is itself an open research question. However, deciding to put these functions outside the purview of search models runs the risk of not being able to model search as it exists in the real world, where these functions are likely interwoven with target guidance and others as part of highly interactive brain circuits. Concluding that these functions are bridges too far is tantamount to concluding that understanding real-world search is outside the scope of search models. It is in the confrontation of how these functions are integrated into an end-to-end attention system that the holes in our understanding of attention will become visible. A related argument is that, yes, these functions are important, but in the interest of keeping focus on "search" one might assume that lower-level processes each perform their function well and pass a "cleaned" version of the visual input up to higher-level processes, such as those engaged in the search for

an object in a scene. But in the absence of knowledge about how the input was cleaned or reconstructed, these models must take as input something other than images. These representations are different from model to model, but the point is that they have become abstracted away from the image input, and therefore cannot be meaningfully compared to image-based models that predict search fixations amidst the noise in the real world. For those models that failed to embrace the techniques needed to use images of complex visual stimuli, they earn in this treatment the delegation of being in the modeling past and no longer representative of current work in predicting search fixations.

3.2 Can your model predict the search for targets of uncertain visual appearance?

Again, your answer to this question may reveal something about your ambitions as a modeler. Promissory notes have abounded in the search literature, and another was a promise sold by modelers that their methods, most of which assumed precise knowledge of the target's appearance, would scale up to the real-world search context where this precise knowledge is never available outside of a laboratory. As shown in Fig. 2A, even car keys, the proverbial target for which unique knowledge might exist from daily exposure, nevertheless appear dramatically different when seen from different perspectives, and depending on how they happened to land when tossed onto the table. The attention literature sometimes refers to target representations as "templates" (Olivers, Peters, Houtkamp, & Roelfsema, 2011), where the assumption is that these templates will rarely match perfectly the target features extracted from a visual input. In the context of search,

Fig. 2 Objects vary in appearance. (A) Examples of variability in perspective and scale. (B) Examples of variability among exemplars of an object category.

there are clear search costs associated with introducing visual mismatch between the target shown at cue and the target as it appears in a search display (Alexander, Nahvi, & Zelinsky, 2019; Vickery, King, & Jiang, 2005; Wolfe, Horowitz, Kenner, Hyle, & Vasan, 2004). However, in the modeling literature, now limited to those that take images as inputs, there is a disconnect from this empirical reality, where they tend to think of the target representation as a literal template that is a pixel-perfect match to the search target. TAM was one of these models. The attention literature, sold on the promissory note, was largely silent on this issue (but see, Zelinsky, Peng, Berg, & Samaras, 2013), perhaps because it recognized the modeling implications of introducing visual uncertainty into the target template.

A related distinction exists between exemplar and categorical search (Fig. 3; Malcolm & Henderson, 2009; Nako, Wu, & Eimer, 2014; Yang & Zelinsky, 2009). Exemplar search is the paradigm in which participants have exact knowledge of what the target will look like in a search display. This target information can be communicated by either showing participants a preview of the target before each search trial, grouping blocks of trials by the same target type, or by using stimuli that are sufficiently simple to obtain precise appearance information based on a verbal or written instruction (e.g., red-vertical bar). By this definition, the majority of studies in the visual search literature have used an exemplar-search paradigm. Models of search have also assumed an exemplar paradigm, where TAM is again an interesting case in point. The input to TAM was two images, one a small image patch of the target and the other a much larger image in which those same target pixels appeared in some search context. From the perspective of the behavioral participant, the small image patch would be presented as a pre-cue designating the target, which would be embedded in the larger search image appearing shortly thereafter. TAM worked by using a bank of linear filters to extract features from the target image, doing the same for the search image, and then taking the dot product between the two. This is a sort of a correlation between

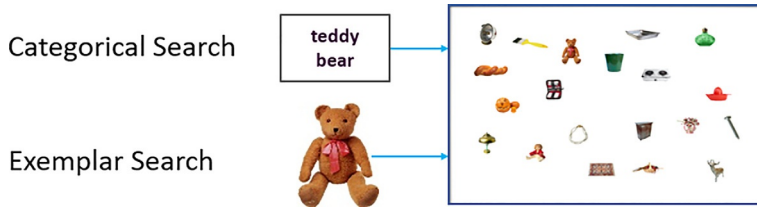

Fig. 3 The designation of a target by category or by exemplar has important implications for models of search.

the target feature vector and an array of like feature vectors computed for each location in an image. This comparison was used to create a target map, which in turn was used to generate a location for the next predicted search fixation. Iterating this process produced for each search image a scanpath of predicted fixation locations by which search brings gaze to the target. However, the features and feature-comparison method used by TAM were not robust to changes in target appearance, such as those that exist between category exemplars, making it largely a model of only exemplar search.

In a categorical search paradigm the search target is designated by category, typically using a text cue preceding each search display (Schmidt & Zelinsky, 2009), or by having a target category designated by instruction and held constant over a block of trials (Yang & Zelinsky, 2009). Unlike exemplar search, in which a person has a very good expectation for how the target will appear, and therefore the specific features to search for, in categorical search this is very often not the case. However, using oculomotor dependent measures it was found that the search for an object category (teddy bears) can indeed be guided, as evidenced by an above-chance direction of initial search saccades to teddy bear targets (Yang & Zelinsky, 2009). This study demonstrated that, even in the absence of scene constraints that might masquerade as a categorical guidance signal (rendered impossible in their study by randomization of target location in object arrays), people can use visual features learned for the target category to guide their gaze to unseen exemplars in a search task. Another study from about the same time showed that the strength of this categorical guidance is proportional to the amount of target-defining information that is provided in the cue (Schmidt & Zelinsky, 2009). For example, stronger categorical guidance was observed when the target cue was "work boot" compared to "footwear" when the search array depicted a work boot. Subsequent work also found that search is guided to distractors that are visually similar to the target category (Alexander & Zelinsky, 2011), for example guidance to a hand fan when searching for a butterfly), that guidance improves with target typicality (Maxfield, Stalder, & Zelinsky, 2014), for example stronger guidance to a dining room chair than a lawn chair when cued with the word "chair", and that guidance becomes weaker as targets climb the category hierarchy (Maxfield & Zelinsky, 2012), for example the guidance to "race car" being stronger than the guidance to "car" and that the guidance to "car" being stronger than the guidance to "vehicle."

The modeling literature on categorical search is not extensive, partly because the problem is so challenging. In addition to the appearance

variability that exists between different views of the same object, there is also appearance variability among the exemplars of an object category, and this variability is often extreme in comparison (see Fig. 2B for exemplars of "waste bins"). Note also that these two sources of variability are not mutually exclusive, depictions of target exemplars in images can also be occluded or viewed from an atypical perspective. So humbling was this problem in its difficulty that researchers in computer vision have been trying to detect objects in scenes for nearly half a century and have only recently achieved what can be called good success (Fan et al., 2020; He, Gkioxari, Dollár, & Girshick, 2017). What makes the problem of categorical search so challenging is that the methods that were developed for exemplar search will not work. Turning to TAM again as an example, the simple V1-like features that it used to represent a target made it very brittle to differences in the target's appearance between cue and test. Although to our knowledge the following experiment was never conducted, one could apply TAM to categorical search by using one of the previously seen exemplars (required because TAM needs to see an image of a target) to create a target map, which can be used to predict fixations in the search task. However, because the success of this prediction would depend on the visual similarity between that specific target exemplar and the one appearing in the search image, the expectation would be that search would be unguided or very weakly guided by such an exemplar-specific model, given our experience with the visual heterogeneity among object exemplars. At least, our level of optimism in reasonable model performance never rose to the level of actually doing this experiment.

Here we argue that the movement from exemplar to categorical search is another sufficiently large milestone in modeling that it justifies carving present from past. The human search experience is that, on any given moment and without need for visual cuing or re-training, we are able to search for hundreds of different target categories. This means that the attention control process must be able to use visual representations of these many different categories of common objects to direct gaze to their likely locations in a visual input upon them becoming a target in a search task. Moreover, the vast and persuasive literature on object-based visual attention and grouping suggests that the visual system also attempts to create objects in the visual input (Scholl, 2001). Moving from exemplar to categorical search is a large step in this context of objects, and one having real cognitive significance. It means that models of goal-directed attention designed to work in the real world can extend to entire object categories, and such movement from exemplars to object categories is likely prerequisite to further movement

into the modeling of object relationships and more human-like semantic structures, as well as the building of perceptual-motor (and robotic) models that achieve a behavioral level of ability to interact with objects. Contemporary models of goal-directed attention need to be engaging the questions of how these object categories are represented and how these representations are used to guide search fixations to targets, and models that do not address this fundamental problem should now be delegated to the past. TAM was progressive for its time in a number of ways. One was its inclusion of a foveated retina in its pipeline of processing, referred to as a *retina transformation*, which transformed the image input to reflect acuity limitations in the visual periphery. TAM was also timely in its use of a relatively high-dimensional feature space to represent visual inputs, was demonstrated to be robust across different types of exemplar targets, and notably was image-based, the other theoretical milestone used in this review. But as discussed, TAM's reliance on precise knowledge of the target's appearance is a fatal weakness, and for this reason it, and models like it (e.g., Pomplun, Reingold, & Shen, 2003), now belong to the past. The problem of categorical search has been a theoretical cliff in the search literature, where models of exemplar search made good progress up to the edge, but then abruptly stopped. Those models of search that embraced the more powerful methods to overcome this obstacle are considered in this treatment to belong in the modeling present.

3.3 Carving past from present

Fig. 4 shows a Venn diagram capturing the aforementioned distinctions in the context of three partially overlapping literatures: *fixation-prediction models*, *image-based models*, and *categorical search models*. Outside the scope of this review are the models addressing goal-directed attention control, but satisfying only two of our three criteria. These *honorable mention* models come in three groups, depending on the missing criterion. The first group consists of models that address the problem of identifying object classes in complex images, but do not attempt to predict the accompanying fixation behavior. In large part, these are the object detection models from the computer vision literature. Object detection is their search task. The task is to locate all the instances of a given category in an image, either by drawing a bounding box around the object or a mask delineating its global contour (Fan et al., 2020; He et al., 2017). Some of these studies even include "attention like" mechanisms (Almeida, Figueiredo, Bernardino, & Santos-Victor, 2017;

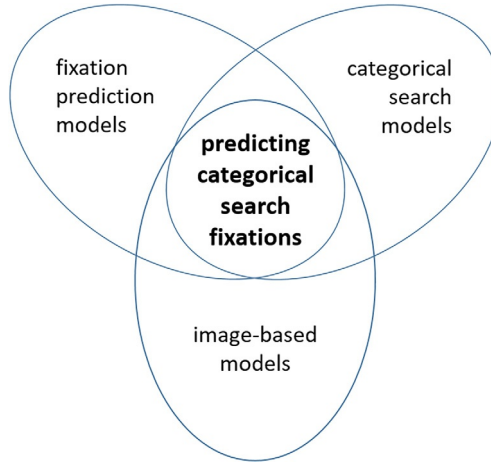

Fig. 4 Venn diagram of modeling milestones, with the intersection of the three defining the scope of this review.

Lan, Ren, Wu, Davis, & Hua, 2020; Shrivastava & Gupta, 2016; Shrivastava, Sukthankar, Malik, & Gupta, 2016), but they really use the concept of attention only as a metaphor for efficient region selection and are not focused on validating model predictions against actual human behavior. In the interest of keeping this review focused on models of fixation prediction, we therefore will not discuss these studies further unless they form the foundations for other fixation prediction models. However, we do consider this fertile ground for advances in search-fixation prediction, if common ground (or application) can be found. Even more deserving of honorable mention are those models that make a more meaningful connection to attention and behavioral vision, some even in the adoption of a foveated input (Akbas & Eckstein, 2017; Butko & Movellan, 2009; Elazary & Itti, 2010; Navalpakkam & Itti, 2005; Yu, Mann, & Gosine, 2011). But given the new bar of fixation prediction for attention models, these models failed to take that final critical step of attempting to predict search fixations. Another group of models that we omit from this review are those that predict fixations in images, but do so in the context of a free-viewing task rather than search (e.g., saliency models; (Cornia, Baraldi, Serra, & Cucchiara, 2018, p. 20; Jia & Bruce, 2020; Jiang, Huang, Duan, & Zhao, 2015; Kummerer, Wallis, Gatys, & Bethge, 2017; Liu & Han, 2018). Special honorable mention goes out to those image-based models that predicted search fixations, but did so only for an exemplar search task and not for categorical search (Hwang, Higgins, & Pomplun, 2009; Zelinsky, 2008).

We already discussed both the fixation-prediction and categorical-search milestones, and our motivations for using them as exclusion criteria. Lastly, there is the group of models that can predict fixations and make categorical distinctions, but require as input either distributions of responses or very simple patterns. This honorable mention group includes the GS models (Wolfe, 1994; Wolfe, 2020; Wolfe et al., 2010; Wolfe et al., 2015; Wolfe & Gancarz, 1997), among others (Bundesen, Vangkilde, & Petersen, 2015). Again, we already discussed our rationale for excluding these models, and the importance of a model being able to scale to images. These models might all be useful in highlighting the many factors that affect search, but their methods are not state-of-the-art and this prevents us from considering them as representative examples of contemporary models of search-fixation prediction in this review.

Models are inherently transient things; they either evolve and turn into different and better models, or they become outdated and are no longer used. This realization helps to sweeten the bitter pill that even some of our most beloved models may belong to the past. However, as distasteful as this realization may be it is important to distinguish between past and present to sustain an active literature on search-fixation modeling. Failing to do so will defocus model comparison efforts and stifle the good work that usually results from this process. It makes no sense to compare contemporary models to the GS models or TAM because these models made assumptions or used information that would undermine any fair comparison. Modeling literatures periodically need to update the playing field and set new rules for the game, and in part this is what we hope to accomplish in this review.

4. The present (∼2010–2020)

Satisfying the three modeling milestones that we discussed is the central region of Fig. 4, which is the focus of this section. These are the models attempting to predict the fixations made during the visual search of an image for a target whose appearance is not precisely known. Because these contemporary models engage the enormously challenging problem of detecting object categories in pixels, it is unsurprising that they borrow heavily from methods for object detection developed in the computer vision literature. Note that this does not mean that we endorse these methods as realistic models for the corresponding behavioral process. Rather, we see them as tools that allow engagement of the question of interest. Tellingly, the computer vision literature, as model rich as it is, has no model comparable to

what behavioral scientists think of as an exemplar search model. Using TAM one last time as an example, the fact that it is image-based does not make it a computer vision model. Indeed, it is not, and for the simple reason that it used methods that would be considered trivial in that literature. If TAM didn't retina transform its image input, the features extracted from a target preview would perfectly match those extracted from the target's location in the search image, resulting in attention moving immediately to the target location almost every time. These perfect matches happen because the target pixels are the same between preview and test, TAM's assumption of perfect knowledge about the target's appearance. With this assumption, even simple V1-like filters can be used to generate a priority map (Rao, Zelinsky, Hayhoe, & Ballard, 2002). But this simplicity comes with the price of very few real-world applications, which is the focus of computer vision. But whereas TAM and models of search behavior have largely failed to advance computer vision methods, the opposite is not true, where computer vision methods have led to significant advances in the prediction of search fixations. Current models of search unashamedly incorporate methods from machine learning to obtain robust feature representations for object goals, and by doing so have achieved some success in predicting fixations in images and under conditions of target-appearance uncertainty. Here we will review the fixation prediction models that have crossed over this bridge to the real world and are now computationally modeling search-fixation behavior in the context of visually complex images.

4.1 Early work using object arrays and pre-deep-network methods

The first image-based fixation-prediction model of categorical search was by (Zhang, Yang, Samaras, & Zelinsky, 2006). Their study used a categorical search paradigm having 6, 13, or 20 images of common objects appearing arrayed into a search display. There was only one target category—teddy bears—and a teddy bear target appeared on half of the trials. For training, color histogram and texture features were extracted from 180 teddy bears and 500 non-bear objects from assorted categories, and AdaBoost (Freund et al., 1996), an algorithm for classification popular then in computer vision, was used to learn a teddy-bear/non-bear classifier. Testing used an additional 180 different teddy bears (the targets in the search arrays), and new non-bear objects served as distractors over the set size conditions. The search array images were input to the teddy bear classifier to create categorical target maps, and these were piped into a version of TAM to

generate sequences of eye movements. These authors compared model fixations to the fixations of people searching for teddy bears in the same object arrays, and found promising agreement using multiple eye movement measures of search efficiency, which included the number of fixations to the target, the cumulative probability of fixating the target as a function of each new search fixation, and the scanpath ratio, which is the ratio of the Euclidean distance between the initial fixation location and the target to the summed Euclidean distances of the eye movements made while searching for the target (i.e., a scanpath ratio of 1 would mean the first eye movement went directly to the target). By predicting this search behavior, this model took the difficult step from exemplar to categorical search, and it did so in the context of visually-complex image stimuli. However, this study was an outlier in the sense that image-based models were at that time still uncommon in the search literature, and much of the recent experimental literature using the categorical search paradigm was yet to be collected. Some time was therefore needed before the modeling literature on categorical search fixations could gain more solid experimental footing, with much of this work appearing in the years surrounding 2010 (Alexander & Zelinsky, 2011; Malcolm & Henderson, 2009; Schmidt & Zelinsky, 2009; Yang & Zelinsky, 2009). Another factor contributing to slow progress was the availability of object datasets at that time for training classifiers. In contrast to the wealth of object datasets that are available today, in 2005 there were very few. Indeed, the reason why teddy bears were used as the target category in the Zhang et al. (2006) study was because one of the authors stumbled upon the *Teddy Bear Encyclopedia* (Cockrill, 1993) one day in a bookstore, and then ripped out its pages and scanned them to create training and testing image datasets.

In another forward-looking study, Ehinger, Hidalgo-Sotelo, Torralba, and Oliva (2009) had participants search for people in 900 images of outdoor scenes, then evaluated several models in their ability to predict categorical guidance in this person search task. This makes their study the first to predict fixation behavior for a target category in natural scenes. A high degree of inter-observer agreement was reported in the search behavior, and the aim of the study was to identify the source of this categorical guidance signal. Three sources of guidance were considered: bottom-up guidance from saliency maps, top-down guidance from target features, and top-down guidance from scene context. To model this behavior the authors modified the contextual guidance model from (Torralba, Oliva, Castelhano, & Henderson, 2006). This model predicted fixations by combining a saliency map, indicating

contrast in local image features, with a bias from a global-scene-context feature. This feature consisted of learned spatial priors reflecting the locations of pedestrians in scenes, which was implemented by a band of heightened priority extending horizontally across an image. The Ehinger et al. (2009) study added to this model a person detector from computer vision (Dalal & Triggs, 2005), which used a texture-based classifier to capture the fact that most pedestrians stand upright and are therefore vertically oriented. It is this addition that makes their model a model of categorical search. In comparing these components, they found that the scene context feature was most important in predicting guidance. This finding is perhaps unsurprising given that people in outdoor scenes often appear on streets or near doors, and that the small size of pedestrians in many images challenged the person detection methods that were available at the time. This study's use of only one target category, and specifically one that is likely not representative of most search targets, also raises concerns about the generalizability of the approach. Nevertheless, this model was an important contribution to the fixation-prediction literature, particularly with respect to the role that scene context can play in categorical guidance.

It wasn't until Zelinsky, Adeli, Peng, & Samaras (2013) that modeling work resumed on categorical search (but see Alexander & Zelinsky, 2011). Similar to Zhang et al. (2006), they extended TAM to the task of categorical search (resisting the urge to call it TAM2) by using a teddy bear classifier to create a categorical target map for the prediction of search fixations. This model used a biologically-plausible and image-based model of object recognition (Serre, Wolf, & Poggio, 2005) to extract features from training images of teddy bears, once again the target category, and random-category non-bear objects. A linear Support Vector Machine (SVM), a classification method that finds a hyperplane best separating labeled data (i.e., teddy bears versus non-bears), was used to train a bear/non-bear classifier. Training and testing used the dataset from Yang and Zelinsky (2009), which had 12 participants search for teddy bears in randomly arranged 6, 13, and 20-object search arrays. Testing consisted of extracting the same features for the objects in the search array, and finding for each object a distance between it and the SVM teddy bear classification boundary. Fig. 5 may be helpful in forming an intuition for the method. The idea is that the features for two object classes, here teddy bears and non-bears, can be separated by a learned boundary in a high-dimensional space (shown for only three dimensions), such that the distance between a given object to this boundary reflects a confidence that the object is a member of the object class.

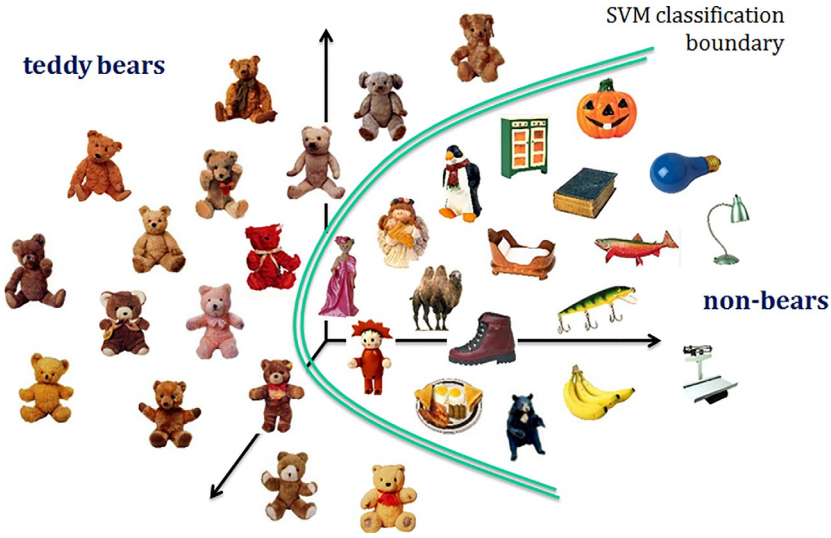

Fig. 5 A simplified illustration of how distance from a learned SVM classification boundary (green line) can be used to prioritize objects for selection in a teddy-bear/non-bear categorical search task.

This is why the pumpkin and doll and black bear objects are positioned close to this boundary, to reflect the fact that they may have visual features in common with teddy bears and that there is lower confidence that these objects are on the correct side of the boundary. Indeed, for the teddy bear dressed in a ball gown, this classification was incorrect. Likewise, the lamp and scale objects are farther from this boundary, reflecting a greater confidence that these objects do not belong in the teddy bear category. From these distances, a categorical target map was created for each search display, and this map of prioritized evidence for the target was piped into TAM to generate saccades. These authors found a near perfect agreement between human behavior and the model with respect to the mean number of fixations made during search across the three set sizes, and the proportion of trials in which initial saccades landed on the target. The Zelinsky, Adeli, et al. (2013) model was in a sense the slow conceptual culmination of the Zhang et al. (2006) model of fixations during categorical search from years before. Related work using this method showed that distance to an SVM classification boundary also predicted the rated typicality of a target exemplar (Maxfield et al., 2014), and could be used to decode the fixations made during target-absent categorical search to classify whether a person was searching for a teddy bear or a butterfly (Zelinsky, Peng, & Samaras, 2013).

In another study from about the same time, Zelinsky, Peng, Berg, et al. (2013) asked whether the categorical guidance signal might largely be object recognition attempted on peripherally viewed inputs. To address this question, these authors trained nine SVM-based teddy-bear detectors, each using different features and methods of the time, on high-resolution images of teddy bear and non-bear objects, a condition that they argued typically exists during recognition only after the high-resolution fovea has moved to an object. They then applied these detectors to teddy bear and non-bear objects that were blurred to approximate viewing in the visual periphery during search, the conditions that exist during target guidance. Using a similar method as in Zelinsky, Adeli, et al. (2013), they found that the most biologically plausible of these teddy bear detectors, and specifically the one which used features extracted by the same object recognition model from the other study, predicted almost perfectly both categorical guidance to the target, measured by the proportion of trials in which the teddy bear was the first fixated object, and the pattern of recognition errors following these initial target fixations. These authors speculated from this finding that categorical search guidance and object recognition may not be substantially different processes, and may in fact be a single process performed on blurred and non-blurred visual inputs, separated by an eye movement. However, and like Zhang et al. (2006) and Ehinger et al. (2009), this modeling work used only one target category and is therefore of untested generalizability.

To study the visual features used to represent target categories, Yu, Maxfield, and Zelinsky (2016) used an unsupervised learning method to extract, directly from images of a category's exemplars, what they referred to as *category-consistent features* (CCFs). They defined CCFs as the features that appear both frequently and consistently across the exemplars of a category, and they extracted CCFs from 4,800 closely-cropped images from 68 common object categories. These 68 categories spanned three hierarchical levels and consisted of 48 subordinate-level categories (e.g., sailboat), which were grouped into 16 basic-level categories (e.g., boat), which were grouped into 4 superordinate-level categories (e.g., vehicle). To identify the CCFs for a category, a scale-invariant texture feature (Lowe, 2004) and a feature consisting of a histogram of color hues (Swain & Ballard, 1991) were extracted from 100 image exemplars making up each of the 48 subordinate-level categories. This was done using yet another image classification method known as Bag-of-Words (BoW; Csurka, Dance, Fan, Willamowski, & Bray, 2004). BoW works by extracting local image features from the exemplars of a category, and then clustering (k-means) these features to obtain a reasonably-sized

vocabulary of "visual words," 1064 in their study. Similar to how text data can be represented using the frequency of words from a vocabulary, representations can be obtained for each exemplar by simply counting how frequently each of these 1064 visual words occurred in a given image, thereby providing a common feature space within which exemplars can be meaningfully compared. Yu et al. (2016) computed BoW histograms for each exemplar of a category, and then determined those features that appeared both frequently and consistently across the category exemplars, a form of signal-to-noise ratio (SNR). The features of an object category having the highest SNRs were selected as the CCFs for that category. The behavioral task was categorical search, where the cue was a category name designating the target in a 6-object search array at a particular hierarchical level. Behavioral responses showed that the time taken by gaze to first land on the target (time-to-target) increased with movement up the hierarchy, replicating the *subordinate-level advantage* in target guidance reported earlier by Maxfield and Zelinsky (2012). The BOW-CCF model captured this subordinate-level advantage by a simple count of the number of CCFs extracted for object categories, grouped at each hierarchical level. This means that more CCFs were extracted from categories at the subordinate level than for categories at either the basic or superordinate levels. The authors interpreted this result as suggesting that the number of CCFs used to represent a visual category is a potential factor affecting the degree of control exerted by that category when it is a target goal.

More recently in the present, Adeli, Vitu, and Zelinsky (2017) developed MASC, an acronym for Model of Attention in the Superior Colliculus. MASC is broadly a model for projecting a cortically-derived priority map onto the surface of the SC, a brain area implicated in eye movement control (Krauzlis, Lovejoy, & Zénon, 2013). Its focus therefore extends beyond search to include free-viewing behavior. MASC's core contribution is that it is a brain-inspired model that takes a priority map as input, be it a saliency map or a target map, and generates sequences of fixations using computations informed by neurophysiological investigations of the SC. In a sense, it is TAM now moved into the brain (see also, Zelinsky, 2012). MASC's pipeline begins with a retina-transformation of the input image, meaning MASC has a foveated retina, and a priority map is computed from this retina-transformed input. This priority map is then projected onto the SC, which is organized into visual and motor spatial maps. In these maps, the priority signals undergo two cascaded stages of Gaussian blurring reflecting population responses known to exist in SC neurons. Winner-take-all is finally used to select the peak in this averaged priority activity, and this location is

selected for the next saccade. Sequences of fixations were produced by iterating this process, inserting an inhibitory spatial tag (Klein, 1988; Mirpour, Arcizet, Ong, & Bisley, 2009) after each movement, and these model fixations were compared to human fixation behavior. To demonstrate the flexibility of their model, the authors reported the success of saliency maps, exemplar target maps, and categorical target maps in predicting the respective behavior from free-viewing, exemplar search, and categorical search tasks, with the lattermost being the reason for its inclusion in this review. In the categorical search task used for behavioral data collection, one target category (from 25 target categories in total) was cued on each trial by name, and this was followed by a search display consisting of randomly-arranged arrays of objects at five levels of set size. The BoW method and texture and color features were used to train 25 target/non-target linear SVM classifiers, one for each of the 25 target categories. Training used only 12 exemplars from each category, none of which were used as targets at test, and 450 random-category non-target objects as negative training samples, also a disjoint set from testing. The authors evaluated MASC on a trial-by-trial basis by computing the feature distances between each search display object and the target's SVM classification boundary, and then converting these distances to probabilities to obtain a categorical target map to make fixation predictions. Search efficiency was measured by summed saccade distance before first target fixation, a measure related to scanpath ratio, where MASC's fixations matched human behavior very well for each of the set sizes tested. In the current context, MASC has the distinction of being the first brain-inspired model of categorical search, even though its categorical target map was simply inherited as a top-down signal.

4.2 More recent work using scenes and deep neural networks to predict categorical search fixations

Even a goal as modest as building fixation prediction models by borrowing already-developed methods from computer vision can be challenging given how rapidly that literature has been changing. Most shocking is the transition that occurred in 2012, the year that AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), an 8-layer convolutional neural network (CNN), decisively beat all competitors in a large-scale object classification competition. This competition was ImageNet; the classification of 1000 different object categories in 1.2 million images (Deng et al., 2009). In the span of one year, the core computer-vision literature abandoned its mainstay methods and shifted almost entirely to deep neural networks (DNNs). This rate of change,

uncommon in behavioral science, meant that features and methods such as color histograms and BoW, along with the models that used them, were flung into the computer vision past, replaced with the more robust features that can be learned by a deep network. Of course, and as already reviewed, categorical search fixations can be predicted without the latest state-of-the-art methods from computer vision, but we would be shortsighted to ignore the wave that is deep learning. The performance benefits from using DNNs are real, and there is no sign of this wave ending soon.

Behavioral scientists interested in visual perception and attention control should care about DNNs because this methodology directly impacts the ability of models in these literatures to use images of scenes as inputs. Using MASC as an example, it is objectively a good model, meaning it is reasonably predictive, demonstrably flexible, and highly explainable. However, there are good reasons why it, and most models before it, used arrays of objects in their predictions of categorical search. For one, these objects, most of which were from the now dated Hemera object collection (Hemera Technologies, 1997), were very well behaved in the sense that objects were photographed from reasonably typical perspectives and were entirely un-occluded. This is not true for visual objects in the wild. Second, using object arrays enables modelers to avoid the problem of object segmentation, itself an open question in computer vision. Hemera objects are again a good example, which come with masks that allow just the object pixels to be placed onto a background, typically a uniform white, to create a search display. This is an important difference from real-world visual scenes, where it is often unclear what should be considered an object or where an object ends and a background begins. Models aimed at predicting goal-directed fixations in realistic contexts must confront these problems or risk not performing very well when tested on scenes. MASC's use of object arrays to test its search predictions therefore exposes a weakness of MASC; the features and methods that it used were able to scale to images of object arrays, but probably not to images of scenes. DNNs are, and will be, an important tool in taking this step because they learn rich representations in their deeper layers that are more abstracted away from the pixel input, arguably how visual-object percepts are more abstracted away from the luminance and color-linked responses of opponent-process retinal ganglion cells.

There are a couple of related literatures using DNNs that will not be covered in this review. One is the excellent literature using DNNs to predict several neural and behavioral responses (e.g., Bao, She, McGill, & Tsao, 2020; Kell, Yamins, Shook, Norman-Haignere, & McDermott, 2018;

Kriegeskorte, 2015; Lotter, Kreiman, & Cox, 2016; Ma & Peters, 2005; Richards et al., 2019; Wang et al., 2018; Yamins & DiCarlo, 2016). These studies do not address goal-directed attention or the problem of predicting fixations, and are therefore well outside our scope. Another is the medical imaging literature that engages the question of fixation prediction during search and uses DNNs to predict cancer diagnoses, but to our knowledge has not yet put these two topics together. However, we give this work more than honorable mention because it comes very close to satisfying the criteria defining our problem, although with caveats. For one, although a radiologist interpreting a mammogram for a nodule is a search task, what this person is actually doing is gathering evidence to reach a decision about the presence of cancer, or its severity. The same is true for the digital pathologist viewing a giga-pixel slide of prostate. This fixation behavior differs from behavior observed in a standard search task, where there is usually a single target located at a single image location. This fixation behavior is also more complex, with visual search often described as being only one of several interacting decision processes (Krupinski et al., 2006; Krupinski, Graham, & Weinstein, 2013; Kundel & Nodine, 1978; Kundel, Nodine, Conant, & Weinstein, 2007). Complicating this goal-directed behavior even further, in mammography it is typical to see multiple views of a case, meaning that the target is distributed across each, and in digital pathology the decision to grade a case as a particular level of cancer is usually reached only after inspection at multiple magnifications. Both of these factors make the definition of the cancer target qualitatively different than that of everyday search targets, such as microwaves. Similarly, although inputs to these models are images, these images are highly specialized and very different in their visual statistics from real-world scenes. The goal of these studies is also not fixation prediction per se, but rather some determination of how close fixation came to target regions during an interpretation and/or how these fixations factored into the medical decision. For example, in a study by Mall, Brennan, and Mello-Thoms (2019) the fixation behavior of 8 radiologists was monitored while they interpreted mammograms from 59 cases of breast cancer. These mammograms were then segmented into three categories of regions: those that were directly fixated, those that were peripherally fixated, and those that were never fixated. A DNN, one pre-trained on ImageNet, was then re-trained on these regions to predict the three categories in a test set of mammograms. The authors found that their model could predict this classification with high accuracy, and concluded that their model enables radiologists to know

whether a particular region in a mammogram is likely to attract foveal or peripheral attention, or no attention at all. This study, and others like it (Brunyé, Mercan, Weaver, & Elmore, 2017), therefore have a very different goal than the other studies of fixation prediction discussed in this review. So, whereas the medical imaging literature does engage the question of categorical search in images, we believe that the nature of the target category, the image, and the task makes these efforts different enough from everyday search that they should be considered addressing different problems.

To our knowledge, the first work using a DNN to predict fixations in a search context was by Wei, Adeli, Zelinsky, Hoai, and Samaras (2016). The study used a VGG16 model (Simonyan & Zisserman, 2014), which is a 16-layer (13 convolutional, 3 fully-connected) deep network that performs well in large-scale object classification, to predict the fixations in images from the POET dataset (Papadopoulos, Clarke, Keller, & Ferrari, 2014). This dataset consists of the fixations made by five people viewing 6270 images, where the participant's task was two-alternative forced-choice, meaning deciding which of two target categories from blocked object-category pairs (e.g., cat or dog; five pairs total) appeared in an image. Because this task is not the same as categorical search (and arguably more similar to an object-classification task for these specific object pairs), and because the focus of the study was a new optimization method in machine learning rather than behavior prediction, we consider this study more squarely belonging to the computer vision literature and will not describe it here in detail. However, the study is notable in that it was one of the earlier attempts to use principles of primate attention to build more efficient computer vision models of object detection. Such "attention" models have since grown in popularity in computer vision (Fu, Zheng, & Mei, 2017; Zamir et al., 2017; Zheng, Fu, Mei, & Luo, 2017; Zoran et al., 2020), making this study a small and long overdue payback to that literature.

The computer vision literature has been evolving at breakneck speed, and it is no coincidence that advances in fixation-prediction models of search have followed shortly behind. In the tradition of first-rate thieves, contemporary models of categorical search have continued seizing upon advances in computer vision methodology with the goal of using these methods to make models of goal-directed attention that are more "end to end," starting with a pixel input and ending with fixation behavior. One of these advances was the use of a convolutional DNN architecture, the CNN architecture used by AlexNet. CNNs are loosely inspired by

the architecture of the brain. They compute features at each network layer by passing filters over an input, a computation similar to what is performed by neurons arranged into retinotopically-organized visual structures. Adeli and Zelinsky (2018) integrated an AlexNet into a model of the broader network of brain structures underlying primate attention control. They did this using the framework of *biased-competition theory* (Desimone & Duncan, 1995; Tsotsos et al., 1995), which captures the currently best understanding of how attention control is mediated by the brain. The theory suggests that a top-down goal state biases a competition among object representations for the selective routing of a visual input for classification. Deep-BCN (Biased-Competition Network) advances this theory by making it explicit as a DNN, and by using it to predict the preferential fixation of categorically-defined search targets. Fig. 6A shows the pipeline of Deep-BCN's neuro-computation, loosely mapped onto processing by early visual structures (V1, V2/V3, and V4), two ventral areas, posterior inferotemporal (PIT) and anterior inferotemporal cortex (AIT), the frontal-eye field (FEF) and the dorso-lateral prefrontal cortex (DLPFC), and the superior colliculus (SC), which recall is the midbrain visuo-motor structure important for eye movement control. Green boxes indicate convolutional DNN layers, yellow

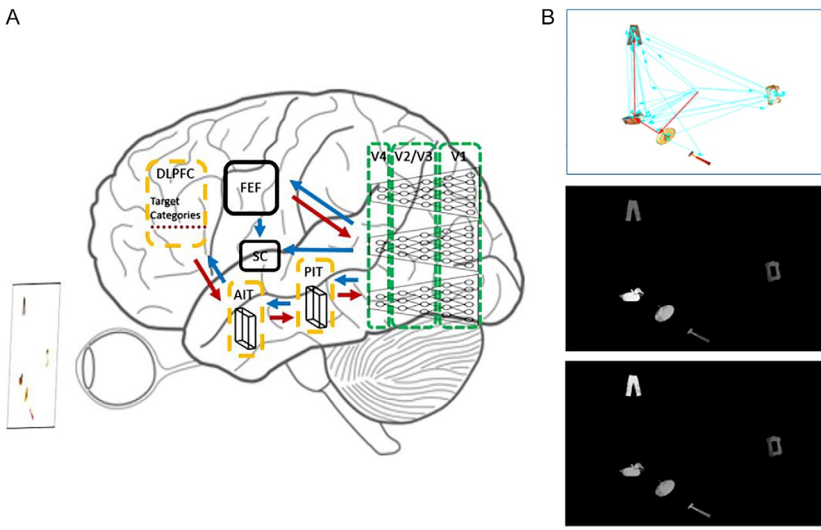

Fig. 6 (A) The pipeline of Deep-BCN. (B) The top panel shows search behavior (model in red, participants in cyan) for the target "pants." The middle panel shows shows the corresponding priority map without the top-down bias, and the bottom panel shows the priority map with the bias for this target category.

boxes indicate fully-connected DNN layers, and black boxes indicate non-DNN processing. Blue and red arrows indicate feedforward and feedback connections, respectively. In this metaphor for the brain, the early visual and ventral components were intended to correspond to the 8 layers of an AlexNet that was pre-trained on ImageNet and fine-tuned on the 25 target-object categories from (Adeli et al., 2017). Deep-BCN's conceptual pipeline assumed that a word cue in a categorical search task activates one of these 25 object categories, and this activity serves as a top-down signal biasing the bottom-up early visual processing in order to prioritize the target location. More specifically, an image of a search display was input to five convolutional layers (green boxes) and three fully-connected layers corresponding to PIT and AIT, brain areas known to be important for object recognition in primates. Importantly, Deep-BCN preserved a coarse retinotopy among units up to its fifth convolutional V4 layer, which is the structure assumed to be biased for the purpose of controlling spatial attention and gaze. This spatial bias was modeled by a feedback connection from the DLPFC layer, and specifically the gradient signal exerted by the learned object category (one of the 25) designated to be the target. The FEF then selected a winner in the biased V4 activity, and this prioritized activation was projected to the SC to generate the search saccade. This final step of generating the eye movement was accomplished by MASC, which was integrated into Deep-BCN's architecture. Fig. 6B (top) shows some of Deep-BCN's eye movements (in red), plotted with representative behavior for a "pants" target on a representative trial. Also shown are two V4 priority maps for this trial, one reflecting purely bottom-up early visual processing (middle) and the other reflecting that processing after it was biased by top-down input (bottom). The authors showed that Deep-BCN's eye movements agreed well with those of people searching the same object arrays for the same categorical targets. They interpreted this agreement as showing how a DNN can model goal-directed attention control in the brain within the context of a biased-competition framework.

In another recent study related to Deep-BCN, Zhang et al. (2018) conducted an impressive data collection effort encompassing three different search contexts. In one task they used arrays of six grayscale objects, one of which was the cued target and the other five of which were exemplars from the five other categories (six categories of targets were tested). They also used a search task in which a target was specified prior to display of a natural image, and a Waldo search task, where the Waldo character was designated as the target once at the start of the experiment, but appeared somewhat differently

in each cluttered and colorfully illustrated search display. Behaviorally, their study largely confirmed what was already known from the previous work on categorical search reviewed here. Specifically, they found that search fixations are guided to categorically defined targets in object arrays (Yang & Zelinsky, 2009), natural scenes (Ehinger et al., 2009), and in a *Where's Waldo* search task (Smith & Henderson, 2011). The greater contribution of this work was their model, which was the first DNN to predict categorical search guidance in the context of scenes. Conceptually, their model works by using a single category exemplar, the one shown to participants at cue, to represent the entire target category. It is therefore categorical search, but instead of using a text cue to designate the target their participants got to see a category exemplar. This is similar to our previous suggestion about how TAM could have been extended to categorical search by using one of its previously viewed target images to extract features. They called this model the invariant visual search network (IVSN). It obtained the activation from the top layer of a pre-trained VGG-16 in response to the target exemplar shown at cue, and used this as a top-down bias for target features in an unseen image. This is similar to the back projection from Deep-BCN's DLPFC layer to its V4 layer, making this method the consensus approach to generating a top-down attention bias in a DNN. Despite their concurrent conception and similarity in design, IVSN and Deep-BCN were developed independently. And there are significant differences. The authors of IVSN interpreted their model as evidence for zero-shot learning, which as the term is used in the computer vision literature means that a model trained to classify exemplars from category **a** is also able to classify exemplars from category **b**, despite never having been trained on **b**. However, this was not entirely true in the case of the Zhang et al. (2018) study, where both participants and IVSN "saw" an image exemplar of the target category, albeit one that differed in appearance from the target in the search displays, and used this target information to guide their search fixations. This type of appearance information is not typically used by computer vision models claiming zero-shot learning, nor was this information available to Deep-BCN or even models dating back to Zhang et al. (2006). This use of target appearance information therefore makes IVSN difficult to situate precisely in the current context. We placed it in the present, both based on its use of DNN methods and its chronology, but, although not nearly as problematic as the methods used by TAM and models like it, IVSN hints at the same problem that caused these models to be delegated to the past.

Brain-inspiration in model design can take many forms, with Deep-BCN and IVSN being two examples. But it is also fair to ask what it was about these models that was particularly brain inspired? Both used pre-trained DNNs to extract the visual features of object categories, and, at least in the case of Deep-BCN, the mapping between network layers and brain areas was intended to illustrate only coarse parallels to the functional connectivity existing between structures in the brain. However, a recent model by Yu, Liu, Samaras, and Zelinsky (2019) was perhaps the most extensive attempt thus far to predict categorical search using a CNN model designed after the brain. This study built on the earlier work by Yu et al. (2016) in two respects. First, it used the same behavioral dataset, which recall was 26 people categorically searching object arrays for each of 68 target categories spanning three levels in a category hierarchy. Second, they borrowed the idea of a category-consistent feature (CCF), which recall is a feature that appears both frequently and consistently across the image exemplars of an object category. One contribution of the Yu et al. (2019) study was the extension of the CCF idea to the features extracted by a DNN. DNNs are powerful because they extract robust feature representations for object categories, more robust compared to previous methods such as BoW, the method used in Yu et al. (2016). Yu et al. (2019) simply replaced the BoW features with the features extracted by a CNN, keeping the CCF extraction algorithm the same. Whereas the BoW-CCF model from Yu et al. (2016) could predict the effect of category hierarchy on search guidance (stronger guidance to targets cued at the subordinate level), using the more powerful CNN-CCF model it was possible to predict search guidance to individual target categories (e.g., stronger guidance to taxis).

A second contribution of the Yu et al. (2019) study was its attempt to design a CNN after the brain structures comprising the primate ventral stream, a goal that motivated the model's name—VsNet. VsNet had five convolutional layers roughly mapping to areas V1, V2, V4, PIT, and AIT, the same ventral areas as in Deep-BCN, followed by two fully-connected layers. For VsNet this mapping meant that each of these layers was engineered to the corresponding brain area in three respects. First, the number of filters in the layer was proportional to the number of neurons in the corresponding structure, based on brain surface area. This was done to have VsNet reflect the brain's distribution of computational resources over the structures in the ventral pathway. Second, the range of filter sizes in each layer was constrained by the range of receptive field sizes for visually-responsive neurons in the corresponding structure.

Third, VsNet implemented bypass connections reflecting known connectivity between areas in the primate ventral stream. VsNet and comparable CNN models were each trained on ImageNet for 2–4 days, then fine-tuned on the SBU-68E dataset. This dataset, expanded from the one used in Yu et al. (2016), has 500 training and 50 validation images for each of 48 object categories. VsNet predicted the time until first target fixation (time-to-target), and it did so both for individual categories as well as the categories grouped by hierarchical level. Moreover, in model comparison these predictions were better than those from other models, despite those models having more trainable convolutional filters. Yu et al. (2019) also probed VsNet to determine the image patches that elicited the maximum responses from filters in its "IT" layers, and observed that many of these images depicted object parts (e.g., a police car siren) that allowed the category to be discriminated from its siblings (e.g., taxis). These authors concluded that CCFs extracted from a brain-inspired CNN, one trained for object classification, were useful in predicting goal-directed attention control. Methods like CCF extraction are potentially useful in adding much-needed explainability to the object representations learned and used by DNNs to predict attention. However, this study focused only on the identification of the ventral stream features that could be used for categorical guidance, and not the source or implementation of the attention-control bias itself. It therefore lacked Deep-BCN's system's level perspective. VsNet's evaluation was also limited only to time-to-target, and not the actual x,y locations of the fixations, or their sequences.

In another study approaching the contemporaneous present, Zelinsky et al. (2019) reported one of the latest efforts to apply DNN methods to predict fixation scanpaths in a categorical search task. The goal of this study was to establish a benchmark for scanpath prediction by applying existing state-of-the-art methods from computer vision to the problem. In an effort purposefully unmotivated by biological plausibility, their approach was to treat the prediction of a search scanpath as a multi-class classification problem. Images were discretized into a 10×16 grid, for problem tractability, and models were trained to classify the grid cell that would be selected (from 160 possible grid locations) for each fixation in the search scanpath. Each fixation prediction in the scanpath also used information from the previous fixations, meaning that these models accumulated foveal information with each search movement. This was done to reflect a memory for fixated objects accumulating over eye movements (Hollingworth & Henderson, 2002), even though the visual details from previous fixations are likely lost once integrated into the new state (consistent with Irwin, 1996). Two approaches were

Fig. 7 Cumulative foveated retina. As fixations accumulate in a scanpath (top, left to right), so too does the high-resolution foveal views obtained at each fixation (bottom, left to right).

considered for representing this accumulated information. The first used a newer generation of CNN, a ResNet-50 (He, Zhang, Ren, & Sun, 2016), which was trained with retina-transformed images of the sort described for TAM. However, and different from previous uses of retina transformation, these movements of a foveated retina accumulated. This is shown in the Fig. 7 example, where each of three shifts of the simulated retina (top) results in the accumulation of high-resolution foveal information from the fixation before (bottom). As the scanpath lengthens, the search image therefore becomes progressively de-blurred. A second approach for representing spatio-temporal search scanpaths used recurrent neural network (RNN) methods, which use a circuit-level memory to process variable-length sequences unfolding over time. The authors reasoned that scanpaths might therefore be a reasonable behavior to predict using this method. Multiple RNN methods were implemented and tested (Cho et al., 2014; Hochreiter & Schmidhuber, 1997), but before diving into their details the authors found that they were not among the most predictive models in their benchmark. Separate models were trained for each target category, and for each model a 6-fixation scanpath was generated for each search image. These scanpaths were compared directly to behavioral scanpaths obtained by people searching the same images for the same target categories. The authors found that all of the models did a good job of predicting the categorical search-fixation scanpaths, relative to a performance ceiling defined by agreement among the behavioral searchers, with the best of these models being a CNN trained with a cumulative foveated retina. But benchmarks of the sort attempted by Zelinsky et al. (2019) are

limited in that they reflect methods existing at only a moment in time, and methods change very quickly in the computer vision literature.

Reaching the current calendar year, Zelinsky et al. (2020) took the novel approach of modeling search-fixation behavior by simply using a DNN to mimic it. We will defer discussion of this technical novelty to the section on the future, where we will make clearer the motivation for this modeling approach. Here we will focus on another methodological contribution of this study, namely the creation of a significant new dataset of categorical search fixations. Details about model training were omitted from the discussion of the Zelinsky et al. (2019) study, most conspicuously where the behavioral fixations came from that were used to train the scanpath-prediction models. These training fixations came from the Microwave-Clock-Search (MCS) dataset, aptly named because it consisted of just two target-object categories—microwaves and clocks. These authors set out to predict goal-directed attention control by training a DNN on previous observations of search-fixation behavior, but before they could do this they needed to collect a lot of categorical search behavior for model training. Two categories were chosen largely arbitrarily from MS COCO (Lin et al., 2014) for behavioral annotation, with two being the minimum number of categories needed to demonstrate different expressions of attention control in the same images. Specifically, 16,184 fixations were collected from people searching for either microwaves or clocks in a training dataset of 4,366 images of scenes, many of which were kitchens because of the microwave category. They used this fixation-labeled dataset with an imitation-learning method from the machine-learning literature called inverse-reinforcement learning (IRL; Abbeel & Ng, 2004) to learn target-specific reward functions and policies for these two target goals. The trained IRL model then used these learned policies to predict the fixations of 60 new behavioral searchers, 30 searching for a clock and the other 30 searching for a microwave, in a disjoint test dataset of kitchen-scene images depicting *both* a microwave and a clock. This design thus perfectly controlled for differences in low-level image contrast, which would be the same regardless of the target. The task was to search for the target category in 80 images, half of which were target-present. As in Zelinsky et al. (2019), this IRL model used a cumulative-foveated image as input and a ResNet-50 backbone to extract object features. Indeed, these two studies differed only with respect to their modeling approach; both used the MCS dataset for model training and testing. They found that the predictions from the IRL model compared favorably to the noise ceiling, based again on agreement in search behavior, and this was true for both fixation-density maps (FDMs) and search-fixation scanpaths. Fig. 8A shows this comparison for

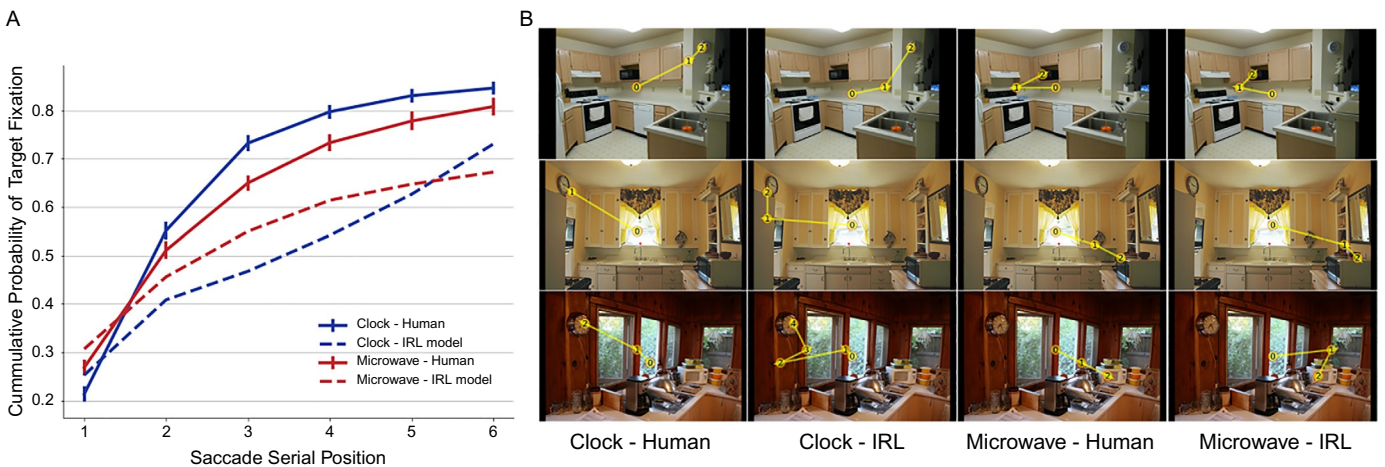

Fig. 8 (A) Cumulative probability of fixating the target as a function of the first six eye movements made during search. (B) Behavioral (human) and model (IRL) scanpaths in clock (left six panels) and microwave (right six panels) search tasks, shown for three representative scenes (top to bottom).

search efficiency, quantified here as the cumulative probability of fixating the target in the first six fixations. This probably increased very rapidly over the first three new fixations for both human searchers (solid lines) and the IRL model (dashed lines), although the behavioral search was still clearly more efficient. But more fundamentally, the IRL model learned target-specific policies that produced target-specific prioritization and guidance in unseen test images, as shown in Fig. 8B in a comparison of representative behavioral and predicted scanpaths. The patterns of fixations made by people and the IRL model clearly depended on whether the target was a microwave or a clock, and this study was the first use of an imitation-learning model to predict this goal-directed control of categorical-search behavior.

5. The future (~2020–2030)

What can we reasonably expect of fixation-prediction models over the next ten years? The past decade has seen good progress in the prediction of goal-directed search fixations, so there is reason to be optimistic and to set our expectations high. But the current state of modeling is also very exploratory, and some time will be needed for the literature to find its footing. The attention modeling literature is experiencing some methodological growing pains. The writing on the wall is clear. If models of attention are to remain relevant and useful, they must be able to predict fixations in realistic contexts. As hopefully captured in this review, the literature has made slow but steady progress from simple search stimuli to images of fully realistic scenes, with arrays of objects serving as a useful stepping-stone along the way. Relatedly, because modeling methods are not yet being developed by and for behavioral scientists, the fixation prediction literature remains tethered to computer vision. While it is unlikely that this will fundamentally change in the foreseeable future, there are things that can be done now to make this tethering less disorienting. As a case in point, the recent and inevitable use of DNNs to predict neural and behavioral responses have almost all used CNNs that were pre-trained to perform object classification. This is because the computer vision literature cares about object classification, so they invested the effort and resources to create datasets and train models that now predict object classification quite respectably. Comparable efforts should be made in the behavioral and neuroscience literature to train DNN models on the specific tasks and behaviors to be predicted. In the current context, this means creating datasets of behavior large enough to train DNN models of search-fixation prediction. DNNs are also a broad class of

models, and its members differ with respect to their architectures and objective functions. There is little consensus among attention and recognition researchers as to which is best for a given task, or what being "best" would even mean. In the absence of data to take a more informed direction, the fixation-prediction literature simply went with the latest pre-trained object classification model that was easily available; Deep-BCN used an AlexNet, IVSN used a VGG-16, and the IRL model used a ResNet-50 backbone. This model evolution was marked by small advances in object classification performance, none of which are likely to be important in the current context. These architecture decisions need to become better motivated by brain or behavior, or both. Relatedly, imitation-learning methods, such as those used in the IRL model, might predict fixations very well, but should prediction be the sole goal of an attention-control model or should a goal also be to understand something about this behavior? What will be the right balance between model performance and interpretability for behavioral attention-control models? If we begin this decade by asking the right questions, hopefully by its end we will emerge from this exploratory phase with a clearer sense of modeling directions. Here we speculate on directions that we see research on attention control taking over the next ten years.

5.1 Search-fixation datasets

DNNs are not magic. To work well, they require training data, and typically a lot. The importance of large-scale datasets for model training is being realized in the attention control literature. This is particularly true in the case of the substantial literature that has developed around the prediction of fixations made during the free-viewing of scenes (https://saliency.tuebingen.ai/). The currently most predictive of these models are DNNs that were pre-trained on SALICON (Jiang et al., 2015), which is a crowd-sourced dataset of 10,000 images that were annotated with ~4,600,000 mouse clicks from people indicating salient image locations. To the extent that SALICON's mouse clicks are comparable to free-viewing fixations, these models are therefore trained on observations of the behavior that they will attempt to predict. Without SALICON, DeepGaze II (Kümmerer et al., 2016) and saliency models like it would not have been possible, and the insights into bottom-up attention control arising from these models might never have occurred.

There is no dataset comparable to SALICON for the training of goal-directed fixation behavior, and certainly nothing approaching this scale when the context is restricted to the fixations made during categorical search.

POET (Papadopoulos et al., 2014) has 6,270 images across 10 target categories, annotated with over 178,000 fixations. However, as discussed in the context of the Wei et al. (2016) study, the participants' task was two-alternative forced-choice object discrimination, which is not the same as object category search. There is also the PET dataset (Gilani et al., 2015), which consists of six categories of target animals in 4,135 images, totaling ~30,000 fixations. The task was categorical search, but it was non-standard in that participants were searching for any exemplar from any of the six target categories, a sort of superordinate categorical search that is known to be poorly guided (Maxfield & Zelinsky, 2012). There were also often multiple exemplars of targets in images, and no images without a target. The MCS dataset from Zelinsky et al. (2019, 2020) consists of ~16,000 fixations on 4,366 images, half of which contained either a microwave (689) or a clock (1,494) target. These fixations were collected using a target-present versus target-absent categorical search task, the standard for the field, but the images varied greatly in search difficulty. This variability reflects a tradeoff in dataset creation between a desire to get as many images as possible for model training (because more is usually better) and a desire to get "good" images that will elicit "good" search fixations. This means fixations that are actually guided to the target categories, which in turn means that the targets should not be too small, located at the center of the image, etc. Although the test images in the MCS dataset were well controlled in this regard, the training images in this dataset, the vast majority, were not. The same is true for all of the other datasets mentioned thus far. Interestingly, the dataset having the most training search fixations (~55,000) for a target category is one of the oldest, the People900 dataset from Ehinger et al. (2009). However, this dataset is comparatively small (912 images, half target-present) and limited to only a single target category, people. Moreover, the people usually appeared in street scenes, making this an atypical target category in that target locations were highly constrained to be on sidewalks or by doors. This means that predictions from search models trained exclusively on this dataset would unlikely generalize to other target categories.

Future directions can sometimes be seen by looking critically at the present, and it was such a realization of the weaknesses in existing datasets of goal-directed attention control that motivated the creation of COCO-Search18. COCO-Search18 is currently the largest dataset of search fixations, an order of magnitude larger than the previously described search datasets. It consists of ~300,000 search fixations collected for 18 target categories, all common objects (microwaves, cups, laptops, etc.), naturally

embedded in a consistent scene context (kitchens, offices, etc.). Search fixations were collected for 6,202 images of scenes selected from MS COCO (Lin et al., 2014). This image selection followed strict inclusion criteria (uncommon in the computer vision literature), and was possible because COCO consists of over 200,000 labeled images and this far exceeds a practical capacity for laboratory-quality eye-movement data collection. Specifically, images were excluded if they depicted: (1) a person or animal, to avoid known attention biases to these object categories (Cerf, Harel, Einhäuser, & Koch, 2008; Judd, Ehinger, Durand, & Torralba, 2009), (2) more than one exemplar of the target, (3) a target that was $<1\%$ or $>10\%$ of the image size, so as not to have the search be too hard or too easy, (4) targets near the center of the image, because that was the starting fixation location, and (5) an odd aspect ratio relative to the display screen, which could distort search behavior. The creators of COCO-Search18 also excluded images if the tightly-cropped target object failed to be confidently detected by a trained object classifier, to exclude images where the target is highly occluded or largely out of frame, and entire categories were excluded if they did not leave at least 100 image exemplars after applying the above exclusion criteria. This latter constraint was introduced because the authors believed that 100 exemplars was the minimum number needed to train a DNN for a specific category. The original sources should be consulted for the full list of criteria, but applying these largely resulted in the selection of 3,101 target-present images from 18 of COCO's 80 object categories. An equal number of target-absent images were selected using additional criteria, the most notable being that images could not contain an exemplar of the target and that the image must depict at least two instances of one of the target's siblings, based on labels in COCO's hierarchical organization. For example, a target-absent image selected for the microwave target category might depict an oven and a refrigerator, both siblings of microwaves under the parent category of appliances. The authors introduced this sibling constraint to discourage searchers from making a target-absent judgment based solely on the context of the scene. Search fixations were collected from 10 participants searching each of these 6,202 images, an effort requiring about 12 hours per participant distributed over the course of 6 sessions, each on a different day. COCO-Search18 belongs in the future because this massive dataset of goal-directed behavior will certainly breed a new generation of models aimed at the prediction of search fixations. Moreover, because COCO-Search18 is now part of the popular MIT/Tuebingen Saliency Benchmark, these models can compete in a

managed competition using withheld testing data. The hope is that such competition will invoke a good-natured adversarial process, thereby accelerating the neurocomputational understanding of attention control. COCO-Search18 can be downloaded from https://saliency.tuebingen.ai/datasets/COCO-Search18/, and references to COCO-Search18 should cite both Yang et al. (2020) and Chen et al. (2020).

5.2 Inverse-reinforcement learning and its applications

Behavioral scientists build and use models for different purposes, but one purpose has always been to predict behavior. Throughout this review we have used the term prediction to mean a generalization from seen to unseen data. This is the definition used in computer vision, where training data typically refers to images for which the model has labels and testing data typically refers to different images for which these labels were not known to the model. Note that behavioral scientists often use the term prediction to refer to model fitting, but these are entirely different endeavors. In model fitting, the model gets to see the test data, and the typical goal is to describe the data pattern using fit parameters. When in doubt, if a study does not mention a training/testing split of their data, which is required only for prediction, chances are the model is data fitting. Returning to the topic of goal-directed attention, essentially two methods for predicting categorical search fixations have been discussed in this review. One is to take a model pre-trained for object classification using object labels, re-train it on the target category with new training labels, and then use this model to predict the fixations made during the search for that target. This was the approach used by most of the models that were discussed. A second method is to do the same, only now also train on the search fixations. Labels can be anything, and just as an image can be labeled with information about an object category, such as "duck," labels can also be behavioral, such as "the x,y location of the third fixation in the search scanpath." This was the approach used by Zelinsky et al. (2020), who trained on observations of microwave and clock search behavior in training images to predict microwave and clock search behavior in unseen test images.

Recall that Zelinsky et al. (2020) used inverse-reinforcement learning (IRL), an imitation learning method from the machine learning literature. The original source should be consulted for details (Ho & Ermon, 2016), but the basic method can be conceptualized as a generator network and a discriminator network that are locked in an adversarial and highly iterative

process, one that is fueled by reward. The generator inputs an image and *generates* a sequence of fixations. These eye movements can be considered fake actions that become paired with a state, mainly features extracted from the input image, to create a specific state-action pairing. However, because there are now behavioral fixations included with the training images, there is also a real state-action pair that was generated by a person searching the same image. The discriminator takes the model-generated and person-generated state-action pairs as input and attempts to *discriminate* the fake from the real. This starts easy, because the generated eye movements are initially random, but each time the discriminator happens to guess wrong, that particular state-action pair from the generator is rewarded, meaning that it will be more likely to be generated in the future. Iterating this process over sufficient training data, the generator becomes increasingly good at fooling the discriminator, which means that it becomes increasingly good at generating human-like fixations that are difficult to discriminate from real. Over training, the model learns a policy for mapping states to actions, such that when it is input a new state (test image) it will be able to predict new actions (search fixations).

Behavioral scientists may want to know about IRL for two reasons. First, simultaneous with learning a policy, the model learns a *reward function*. The reward function tells you how much total reward would be expected from making a sequence of actions, in this case fixations given an image and a learned target category. For the IRL model, the attention priority map is neither a saliency map nor a target map, but rather a *reward map* indicating where the next fixation should be directed, all with the goal of maximizing with each fixation the total expected reward. Our premise is that reward drives search behavior, as it does many others (Anderson, 2013), meaning that people select fixation locations during search in the pursuit of a reward that will be derived upon finding the target. This is that small jolt of dopamine delivered by your reward system the moment your eyes finally land on your keys after a frantic search when rushing to catch a train, and imagine how more important efficient target acquisition would have been throughout our evolution. All behavioral scientists are on board with the idea that reward is an important force in shaping behavior, but is this Skinnerian sense of reward the same reward that is learned by the IRL model? As with many good questions, the answer is "yes" and "no." No, in the sense that what the model is reinforced for is imitating fixation behavior, which is not the same as finding keys. But yes, in the sense that the behavior that is imitated is categorical search, which by our premise is

driven by the expectation of reward by the searcher. This is why the IRL model from Zelinsky et al. (2020), although not rewarded directly for target fixation, nevertheless indirectly recovered separate reward functions for predicting microwave and clock search. A second reason why behavioral scientists should know about IRL is that it is a powerful tool for predicting behavior in a task, perhaps because of its grounding in reward, and this creates opportunities for new behaviorally-engineered applications. Very speculatively, with IRL it might be possible to obtain reward functions for populations that have malfunctioning reward systems or specialized reward histories, and make testable predictions for how these people prioritize visual inputs. For example, how would reward functions recovered for a typically-developing population of children compare to ones from children who are autistic, or how would the reward functions from people with well-functioning reward systems compare to those from people who are depressed, phobic, or who suffer from any anxiety disorder or post-traumatic stress disorder that is believed to be expressed in changes of fixation behavior (e.g., Armstrong & Olatunji, 2012; Cisler & Koster, 2010; Papagiannopoulou, Chitty, Hermens, Hickie, & Lagopoulos, 2014)? There might also be applications in human-computer interactive systems that can anticipate a person's intent and render assistance, and clear applications in education, learning, and training environments. Because IRL imitates behavior, differences can be quantified between people who have learned a highly specialized visual skill and those who have not. How does the fixation behavior of a digital pathology resident student still learning how to detect grade-five cancer in a slide of prostate compare to the fixation behavior of a more advanced genitourinary pathologist at a hospital who regularly sees such cases clinically, and can the quantification of this comparison lead to better tools for training this student? Its potential to address important problems, ranging from diagnosing the severity of attention-related disorders to evaluating the achievement of expertise in some visual task, is why we see IRL as a future direction for the application of fixation-prediction models.

So what has been done so far? The current state-of-the art in predicting the fixations made during categorical search is a study by Yang et al. (2020), which trained IRL models using COCO-Search18. IRL was also used with the MCS dataset (Zelinsky et al., 2020), but this dataset had only microwave and clock target categories. COCO-Search18 extends this to 18 types of target-object goals, thereby enabling the model to learn from a richer diversity of target categories and far more search fixations. The IRL model in this study learned a reward function and policy from people searching

COCO-Search18, and it did this using a cumulative foveated retina similar to what was done in Zelinsky et al. (2019, 2020) to capture the dynamic change in visual state that occurs with each behavioral search fixation. However, the greater technical novelty of this work was in a modeling of the searcher's changing state of knowledge about the objects in the image, what they referred to as Dynamic Contextual Beliefs (DCB). The DCB representation captures a person's "what" and "where" understanding of a scene as it evolves during the course of search. For example, given a particular type of kitchen scene it is highly likely that a microwave oven will appear above the stove. This means that if a stove, a usually large object, is detected in peripheral vision, then a model trained on a DCB representation might attach higher probability to the blurred blob above it being a microwave. The theoretical assumption was that a visual input is parsed into contextual belief maps based on all the visual categories in a person's knowledge structure. Belief maps were approximated using a panoptic segmentation method (Kirillov, Girshick, He, & Dollár, 2019; Kirillov, He, Girshick, Rother, & Dollár, 2019), which segmented an input image into any of 80 "object" or 54 "stuff" categories from COCO (Caesar, Uijlings, & Ferrari, 2018). Their IRL model therefore used information, not just from a changing visual state filtered through a foveated retina, but also from a 134-dimensional contextual belief map that changed with each fixation, all for the purpose of better imitating the behavioral search fixations. After training, this model learned a policy and reward function for pairing these perceptually more abstract states with sequences of search fixations, and these functions were used to predict the search fixations in the testing dataset. These predictions happened on a fixation-by-fixation basis, enabling a comparison between the cumulative probability of target fixation in the first six eye movements made during search, among other search efficiency measures. In extensive model comparison, the authors' IRL model best predicted search behavior on this gold-standard measure of target guidance, even when compared against other imitation learning methods. Similarly good performance was shown across a range of fixation measures and metrics. They speculated that this greater predictive success was due to an alignment in the optimization process performed during search; both the IRL model and people make fixations that maximize the accumulation of total expected reward. But perhaps the more enduring contribution of this future-pointing study was the fact that these reward functions included a broader object context from the DCB representation. Note that this can result in objects being prioritized despite looking very little like the target. As shown in Fig. 9, the stove and sink are prioritized

Fig. 9 Prioritization by object context. Left: search image, with a microwave designated as the target. Right: the reward map generated for this image by the IRL model. Note that the stove and sink are prioritized because these objects carry information about the location of a microwave.

despite these objects being very different in appearance from the microwave, which was the target. This broader prioritization was interpreted as the attachment of reward to an object context, enabling even non-target objects to assist in the guidance of search fixation. The Yang et al. (2020) study, in addition to offering the currently most predictive model of categorical search in scenes, is therefore also the latest attempt to learn a visual context for target guidance, significantly extending the seminal work on this topic by Ehinger et al. (2009). We see the potential for learning contextual guidance being another important future direction for attention control modeling.

5.3 Building more brain-inspired models

It took two tries, but neural networks, now that they are deep, are here for the foreseeable future. The accelerating use of DNN models in the visual attention and recognition literatures gives reason for optimism about what can be learned over the next decade. It means that researchers are converging on a common methodology, and will therefore be able to speak the same language in their theories. This quick pace is also good because it means that the modeling literature should evolve quickly. But in the frenzy of modeling studies that are sure to come it will be good to periodically ask what exactly has been learned. The last decade has made good progress in the use of DNNs to predict fixations, goal-directed and otherwise, but has anything fundamental been learned about attention from all this modeling? In the case of categorical search, one might argue that it is now possible to learn the feature representations that guide attention to a category of target, which *is* something, but the representations learned by the models discussed in this review exist in such a high-dimensional space that it becomes unclear how this modeling advance translates into a greater understanding of attention

control. We believe the rapid embrace of DNNs by behavioral scientists is a good thing, but the attention literature should be aware of the implications of entering into this pact. The features of an object category learned by a DNN will not be comprehensible in the same way as the features of a red vertical bar. Although very promising methods exist for forming more compact visual representations of objects (Sabour, Frosst, & Hinton, 2017), even these will unlikely be verbalizable. Indeed, even identifying the features of a DNN that are used for attention control is challenging, although techniques are available to do this (e.g., as in Yu et al., 2019). If the attention literature is sincere in its adoption of DNN methodology, it must therefore be prepared to abandon, completely and forever, the hope of understanding the representations of attention control using a language of simple visual features. This shift in expectation is long overdue, and reflects the attention literature being misled into believing that these representations would be this simple—that this promise would be delivered—when in fact there was every reason to believe exactly the opposite. So, whereas a search future focused on application may find the prediction of goal-directed fixations very valuable, the literature attempting to understand something more basic about attention control is right to question the value of models that learn feature representations of such complexity that they defy understanding (although it is not obvious what "understanding" even means in this context). The computer vision literature recognizes this problem, and is currently valuing models in proportion to their degree of explainability. The attention literature will need to do the same, at least with respect to understanding the basic nature, source, and destination of attention biases in a model.

It is as yet unclear the degree that the behavioral literature will remain yoked to computer vision methods, assuming a continued embrace of DNNs by attention and recognition researchers. Some degree of connection is healthy, but these literatures must better differentiate themselves if their focus is to remain on questions pertaining to brain and behavior. This means that models must be trained and evaluated on behavioral and neural responses, and promising architectures and learning rules must not be excluded if they fail to win large-scale object-classification competitions. One way that the behavioral literature is starting to differentiate itself from computer vision is in the design of brain-inspired models. This has already started to occur in models of object recognition (e.g., Dapello et al., 2020), and also for models of attention that were not focused on fixation prediction (Lindsay, Rubin, & Miller, 2019). With respect to the problem at hand, the direction suggested by Yu et al. (2019) is a start, but far more work is needed. Finding the most useful

sources of brain inspiration to build into a DNN is a challenging problem. Using the Yu et al. (2019) study as an example, its brain-engineering focused on kernel sizes and skip connections, but obviously absent was a role for the recurrent connections that are known to exist in the brain (Gilbert & Li, 2013). Including these temporal interactions in some future version of VsNet is an attainable goal over the next years. Relatedly, another form of brain inspiration will be the design of fixation-prediction models aimed at capturing the functional connectivity between brain areas in the broader attention network. Deep-BCN is an early example of this, which included a recurrent process in the form of a top-down bias that it suggested might be exerted on a ventral pathway structure for the purpose of controlling search fixations, but this bias was implemented in a static way that lacked real brain inspiration. DNN models of attention will need to become more dynamic in order to better reflect the flow of processing and information occurring throughout the brain. Even more challenging is when brain inspiration extends beyond model architecture and into state representations and learning rules. The study by Yang et al. (2020) showed the potential of exploring higher-level state representations, and Chen et al. (2020) started an effort to inform how different types of foveated retinas might be implemented and used as state representations in the training of DNN models of search-fixation prediction. This work should continue. Minimally, we should hope to find consensus in the near future on how differences between central and peripheral visual inputs should be represented for model training, and the fact that multiple labs have begun to explore this issue is a promising sign that progress will be made on this front (Deza & Konkle, 2020). The attention-modeling literature must also decide on how to integrate into DNNs the reward-based circuitry known to be used by the brain, and think carefully about its assumptions of supervision during training, particularly with regard to what labels are used and where they come from. The literature should brace itself for the possibility that supervised, reward-based, and unsupervised learning might all be happening in the brain at the same time in different architectures, making even a roughly complete understanding of the primate attention network a question that will likely remain open throughout the next decade.

5.4 Understanding the attention-control network

What will models of spatial attention look like moving forward? In the near future they will likely continue the trend of using DNNs to approximate ventral visual processing in the primate brain, as already done by Deep-BCN,

VsNet, and others. This visual component of a model is important because here is where feature representations are learned that give a model its robustness to variability in object appearance, essential to categorical search. There is no going back from there. As for what these models will be able to predict, one can only speculate, and hope. Here we sketch our hope for what future attention models will become. Fixation-prediction models will need to acknowledge more consistently the fact that the visual input is from a foveated retina, without which eye movements would be unnecessary. But aside from this anatomical reality, each movement of a foveated retina changes the visual input in nontrivial respects, and it is reasonable to assume that these retina-transformed inputs affect the selection of image locations to fixate. This is especially true for goal-directed allocations of attention, where these goals are often objects appearing blurred due to peripheral viewing. Objects that can be recognized accurately in cropped, high-resolution image patches, might not be accurately recognized, or correctly recognized but at a lower level of confidence, when seen in the blurred visual periphery. It therefore follows that eye movements during search are made to increase the confidence of goal decisions by bringing the high-resolution fovea to new locations in the visual field. Each fixation therefore obtains a high-resolution glimpse of any object appearing at its location, cropped by the dimensions of the fovea, followed by another search saccade if this glimpse fails to reveal sufficient evidence for the target decision. This is how TAM worked, and its articulation of this confidence-based oculomotor dynamic during search is perhaps the model's greatest lasting legacy. However, with newer methods researchers can drill deeper into this theoretical bedrock, and there are several valuable directions to go. One direction was hinted at in the Yang et al. (2020) study, where priority was computed at each fixation based on dynamic contextual beliefs. Models of the future will need to explore state representations consisting of spatially-localized hypotheses for the objects existing at different locations in a retina-transformed image, and how these hypotheses dynamically change as information accumulates from earlier fixations over the visual input. A model that better captures this changing hypothesis structure, and integrates it with a learned scene and object context, will hopefully exist before the decade's end. Another fruitful direction, related but also parallel to the first, is to form stronger connections between goal-directed visual attention and other perceptual processes, with a good candidate for this being visual object recognition.

Any understanding of visual attention will be incomplete without considering its interaction with object recognition and its precursor processes,

and models will play an important role in understanding this attention-recognition system in the brain. There is already an emerging literature on unified attention-recognition models (Adeli & Zelinsky, 2018; Lindsay & Miller, 2018) and we anticipate that this literature will grow rapidly. We see the relationship between recognition and goal-directed attention control being a promising future direction for this growth, and in particular a better integration of recognition into the changing state of fixations in a search scanpath. We also anticipate that an important bridge in this unification effort will be the inclusion of visual grouping and figure-ground segmentation processes into models. The segmentation of an object from a complex background is not an all-or-none thing; some can be more exact than others. In computer vision there is a field known as object segmentation, the task of segmenting an instance of an object category from the rest of the image, where the assumption is that there will usually be some error with respect to the actual global contour for that object instance. In that literature there is even the concept of an *object proposal*, a region in an image that is likely to be an object, although it is not known what kind of object. These methods are ripe for harvest by behavioral scientists, and collectively with goal-directed attention and recognition one can start to see the making of a unified system. The spatial exactitude of object segments and proposals will very likely depend on how far in the visual periphery these exist. Ones forming nearer to high-acuity central vision will obviously better delineate an object compared to those extracted from progressively blurred eccentricities in the visual periphery. This inexact object segmentation will in turn inject uncertainty in the scene's dynamic hypothesis structure, ultimately lessening the potential for attention control. Under this view, the consequence of an eye movement is to improve object recognition confidence by shifting the high-resolution fovea to an image location, but now with a focus on the role that segmentation plays in this process. The fixation of an object focuses figure-ground segmentation processes on a relatively small region of space, thereby improving the shape estimation of the centrally-viewed object and increasing its recognition confidence, and potentially its accuracy. Assuming that the target object must be recognized at some threshold level of confidence before committing to the manual judgment, then attention control during search may largely be the process of selecting image locations for fixation that will increase target recognition confidence. However, such intuitive relationships mask the complexities involved in simultaneously incorporating these attention, segmentation, and recognition processes into the dynamics of a single DNN model. To get a handle on this problem, it may even be

prudent in the near term to focus on simpler models that address pairwise combinations of these processes. For example, can models be developed that learn feature representations that predict the relationship between figure-ground segmentation and attention, such that poorer target segmentation in the visual periphery leads to a greater number of eye movements needed to fixate the target? Can other models be developed to show that the features used by attention to bias gaze shifts to the target are the same features used to recognize that object upon its fixation? This was the suggestion from Zelinsky, Peng, Berg, et al. (2013), who indeed advocated for researchers to adopt the categorical search paradigm specifically to study the relationship between goal-directed attention and object recognition. Categorical search was a paradigm created in part specifically to study the attention-recognition interaction, but even more convincing would be to have models that generalize across different responses using that paradigm. Can the feature representations learned from a model predict both the fixation behavior of a monkey searching for a categorical target and the activity of target-tuned inferotemporal neurons at each of the fixations (similar to Sheinberg & Logothetis, 2001)? Such converging evidence, either across responses or across paradigms (e.g., model predictions generalizing from a search task to a segmentation task), would begin to end the purely exploratory phase that the modeling literature is in now, and start to settle on the architectures and learning rules that will define the future modeling of visual perception.

The above-sketched relationship between visual spatial attention, recognition, and segmentation is, when put in the context of fixation behavior, an observable physical implementation of biased-competition theory (Desimone & Duncan, 1995; Tsotsos et al., 1995), and specifically what we refer to as *fixation-selective routing*. We suggest that fixation-selective routing is consistent with the core principles of biased competition theory, only at the scale of a competition between object beliefs for overt attention and confident recognition. The function served by search fixations under the sketched view is very similar to the hypothesized tuning function performed by selective routing under biased-competition theory. Broadly summarized, that theory proposed that attention control is the process of biasing a competition between feature representations of visual inputs for the purpose of selectively routing the inputs from only one to higher processing levels, such as those involved with object recognition. We contend that the fixations made in the course of search are performing this exact function. The selection of an object for fixation is itself a competition between simultaneous object representations in an image, with the winner

of this competition getting to have its shape features, along with all the rest, better tuned due to their extraction from high-resolution central vision. As a consequence of this fixation, cleaner visual inputs are selectively routed to higher processing areas, resulting in a more confident object classification. We believe that this fixation-selective routing happens in two feedforward stages. First, object-based grouping processes are used to extract shape precursors to objects. Related work exists on the extraction of proto-objects from images (Yu, Samaras, & Zelinsky, 2014), but significant study is needed on how best to model these object-like entities. On this point, we see capsule networks (Hinton, Sabour, & Frosst, 2018; Sabour et al., 2017) as showing promise in their ability to form stable, yet still fleeting, representations of shape. Second, these proto-objects are then routed to the object classification process that will be used to control behavior, which can be either the manual button press terminating a search or the eye movement to the next most probable location of the target object. Grounding fixation-prediction models in biased-competition theory is useful, not only because it is a widely accepted framework that can serve as a consensus theoretical platform for modelers, but also because this theory has emerged as modern doctrine among neurocomputational models of attention (Hamker, 2006; Tsotsos, 2011), increasing the likelihood that model predictions will actually become hypotheses that are tested using neuroscience techniques. Will such a unified model be able to predict grouping, fixation, and recognition behavior? That is the final hope, but we suspect that realizing this model will take some work, and may even require veering for a period of time back to using simple stimuli for model development. Here we see arrays of alphanumeric characters as being particularly valuable, given that they are highly familiar learned visual categories and have been used extensively in the early search and recognition literatures. An opportunity therefore exists for some very old research to inform new future directions. But more crucially, alphanumeric stimuli engage core grouping, spatial attention, and recognition processes, albeit each in highly simplified domains. We suspect this simplicity might be needed to train the first DNN models of the attention-recognition system. If such a detour occurs, the end goal should be to scale these models back up to real-world visual complexity as soon as possible, because only at this level might fundamental limitations become apparent. At the time of this writing, realizing such a model seems like a milestone that a future author reviewing the next generation of goal-directed fixation-prediction models may use to separate past from present. This was, is, and will be an exciting time to model search fixations.

References

Abbeel, P., & Ng, A. Y. (2004). In *Apprenticeship learning via inverse reinforcement learning* (p. 1). Proceedings of the twenty-first international conference on machine learning.

Adeli, H., Vitu, F., & Zelinsky, G. J. (2017). A model of the superior colliculus predicts fixation locations during scene viewing and visual search. *Journal of Neuroscience*, *37*(6), 1453–1467.

Adeli, H., & Zelinsky, G. (2018). In *Deep-BCN: Deep networks meet biased competition to create a brain-inspired model of attention control* (pp. 1932–1942). Proceedings of the IEEE conference on computer vision and pattern recognition workshops.

Akbas, E., & Eckstein, M. P. (2017). Object detection through search with a foveated visual system. *PLoS Computational Biology*, *13*(10), e1005743.

Alexander, R. G., Nahvi, R. J., & Zelinsky, G. J. (2019). Specifying the precision of guiding features for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(9), 1248.

Alexander, R. G., & Zelinsky, G. J. (2011). Visual similarity effects in categorical search. *Journal of Vision*, *11*(8), 1–15.

Allport, D. A. (1980). Attention and performance. *Cognitive Psychology: New Directions*, *1*, 12–153.

Almeida, A. F., Figueiredo, R., Bernardino, A., & Santos-Victor, J. (2017). In *Deep networks for human visual attention: A hybrid model using foveal vision*(pp. 117–128). Iberian Robotics conference, Springer.

Anderson, B. A. (2013). A value-driven mechanism of attentional selection. *Journal of Vision*, *13*(3), 1–16.

Armstrong, T., & Olatunji, B. O. (2012). Eye tracking of attention in the affective disorders: A meta-analytic review and synthesis. *Clinical Psychology Review*, *32*(8), 704–723.

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, 1–6.

Borji, A. (2019). Saliency prediction in the deep learning era: Successes and limitations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Broadbent, D. E. (1957). A mechanical model for human attention and immediate memory. *Psychological Review*, *64*(3), 205.

Brunyé, T. T., Mercan, E., Weaver, D. L., & Elmore, J. G. (2017). Accuracy is in the eyes of the pathologist: The visual interpretive process and diagnostic accuracy with digital whole slide images. *Journal of Biomedical Informatics*, *66*, 171–179.

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, *97*(4), 523.

Bundesen, C., Vangkilde, S., & Petersen, A. (2015). Recent developments in a computational theory of visual attention (TVA). *Vision Research*, *116*, 210–218.

Butko, N. J., & Movellan, J. R. (2009). In *Optimal scanning for faster object detection* (pp. 2751–2758). 2009IEEE conference on computer vision and pattern recognition, IEEE.

Caesar, H., Uijlings, J., & Ferrari, V. (2018). In *Coco-stuff: Thing and stuff classes in context* (pp. 1209–1218). Proceedings of the IEEE conference on computer vision and pattern recognition.

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. *Advances in Neural Information Processing Systems*, 241–248.

Chen, Y., Yang, Z., Ahn, S., Samaras, D., Hoai, M., & Zelinsky, G. (2020). COCO-Search18: A dataset for predicting goal-directed attention control. *bioRxiv*.

Chen, X., & Zelinsky, G. J. (2006). Real-world visual search is dominated by top-down guidance. *Vision Research*, *46*(24), 4118–4133.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *ArXiv Preprint ArXiv*, *1406*, 1078.

Cisler, J. M., & Koster, E. H. (2010). Mechanisms of attentional biases towards threat in anxiety disorders: An integrative review. *Clinical Psychology Review*, *30*(2), 203–216.
Clifton, C., Jr., Ferreira, F., Henderson, J. M., Inhoff, A. W., Liversedge, S. P., Reichle, E. D., et al. (2016). Eye movements in reading and information processing: Keith Rayner's 40 year legacy. *Journal of Memory and Language*, *86*, 1–19.
Cockrill, P. (1993). *The teddy bear encyclopedia*. Barnes & Noble Books.
Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R. (2018). Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, *27*(10), 5142–5154.
Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, ECCV*, 1(1–22), 1–2, Prague.
Dalal, N., & Triggs, B. (2005). In *Histograms of oriented gradients for human detection* (pp. 886–893). 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), 1, IEEE.
Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D. D., & DiCarlo, J. J. (2020). Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations. *BioRxiv*.
Deco, G., & Rolls, E. T. (2004). A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, *44*(6), 621–642.
Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009IEEE Conference on Computer Vision and Pattern Recognition*, 248–255, IEEE.
Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193–222.
Deza, A., & Konkle, T. (2020). Emergent properties of foveated perceptual systems. *ArXiv Preprint ArXiv*, *2006*07991.
Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, *17*(6–7), 945–978.
Elazary, L., & Itti, L. (2010). A Bayesian model for efficient visual search and recognition. *Vision Research*, *50*(14), 1338–1352.
Fan, D.-P., Ji, G.-P., Sun, G., Cheng, M.-M., Shen, J., & Shao, L. (2020). Camouflaged object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2777–2787.
Findlay, J. M. (2004). Eye scanning and visual search. *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*, *134*.
Findlay, J. M. (2005). Covert attention and saccadic eye movements. In *Neurobiology of attention* (pp. 114–116). Elsevier.
Findlay, J., & Gilchrist, I. (2003). *Active vision: The psychology of looking and seeing*. New York, NY, US: Oxford University Press.
Foulsham, T., Walker, E., & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, *51*(17), 1920–1931.
Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. *Icml*, 96, 148–156, Citeseer.
Fu, J., Zheng, H., & Mei, T. (2017). Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4438–4446.
Gilani, S. O., Subramanian, R., Yan, Y., Melcher, D., Sebe, N., & Winkler, S. (2015). PET: An eye-tracking dataset for animal-centric Pascal object classes. 1–6. IEEE.
Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*(5), 350–363.

Hamker, F. H. (2004). A dynamic model of how feature cues guide spatial attention. *Vision Research*, *44*(5), 501–521.
Hamker, F. H. (2005). The reentry hypothesis: The putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement. *Cerebral Cortex*, *15*(4), 431–447.
Hamker, F. H. (2006). Modeling feature-based attention as an active top-down inference process. *BioSystems*, *86*(1–3), 91–99.
He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). In *Mask r-cnn* (pp. 2961–2969). Proceedings of the IEEE international conference on computer vision.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). In *Deep residual learning for image recognition* (pp. 770–778). Proceedings of the IEEE conference on computer vision and pattern recognition.
Hemera Technologies. (1997). *Hemera photo-objects: 50,000 Premium image collection [document Électronique]*. Hemera Technologies.
Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498–504.
Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In *Eye movements*. (pp. 537–III). Elsevier.
Henderson, J. M., Shinkareva, S. V., Wang, J., Luke, S. G., & Olejarczyk, J. (2013). Predicting cognitive state from eye movements. *PloS One*, *8*(5), e64937.
Hinton, G. E., Sabour, S., & Frosst, N. (2018). In *Matrix capsules with EM routing*. 6th International conference on learning representations, ICLR.
Ho, J., & Ermon, S. (2016). Generative adversarial imitation learning. *Advances in Neural Information Processing Systems*, 4565–4573.
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.
Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(1), 113.
Hunt, A. R., Reuther, J., Hilchey, M. D., & Klein, R. M. (2019). The relationship between spatial attention and eye movements. *Processes of Visuospatial Attention and Working Memory*, 255–278.
Hwang, A. D., Higgins, E. C., & Pomplun, M. (2009). A model of top-down attentional control during visual search in complex scenes. *Journal of Vision*, *9*(5), 1–18.
Irwin, D. E. (1996). Integrating information across saccadic eye movements. *Current Directions in Psychological Science*, *5*(3), 94–100.
Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203.
Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.
Itti, L., Rees, G., & Tsotsos, J. K. (2005). *Neurobiology of attention*. Elsevier.
Jia, S., & Bruce, N. D. (2020). Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing*, *103887*.
Jiang, M., Huang, S., Duan, J., & Zhao, Q. (2015). In *Salicon: Saliency in context* (pp. 1072–1080). Proceedings of the IEEE conference on computer vision and pattern recognition.
Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). In *Learning to predict where humans look* (pp. 2106–2113). 2009IEEE 12th international conference on computer vision, IEEE.
Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630–644.

Kirillov, A., Girshick, R., He, K., & Dollár, P. (2019). In *Panoptic feature pyramid networks* (pp. 6399–6408). Proceedings of the IEEE conference on computer vision and pattern recognition.

Kirillov, A., He, K., Girshick, R., Rother, C., & Dollár, P. (2019). In *Panoptic segmentation* (pp. 9404–9413). Proceedings of the IEEE conference on computer vision and pattern recognition.

Klein, R. (1988). Inhibitory tagging system facilitates visual search. *Nature*, *334*(6181), 430–431.

Koehler, K., Guo, F., Zhang, S., & Eckstein, M. P. (2014). What do saliency models predict? *Journal of Vision*, *14*(3), 1–27.

Krauzlis, R. J., Lovejoy, L. P., & Zénon, A. (2013). Superior colliculus and visual spatial attention. *Annual Review of Neuroscience*, *36*.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks*. 1097–1105.

Krupinski, E. A., Graham, A. R., & Weinstein, R. S. (2013). Characterizing the development of visual search expertise in pathology residents viewing whole slide images. *Human Pathology*, *44*(3), 357–364.

Krupinski, E. A., Tillack, A. A., Richter, L., Henderson, J. T., Bhattacharyya, A. K., Scott, K. M., et al. (2006). Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience. *Human Pathology*, *37*(12), 1543–1556.

Kümmerer, M., Wallis, T. S., & Bethge, M. (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. *ArXiv Preprint ArXiv*, *1610*01563.

Kummerer, M., Wallis, T. S., Gatys, L. A., & Bethge, M. (2017). In *Understanding low-and high-level contributions to fixation prediction* (pp. 4789–4798). Proceedings of the IEEE international conference on computer vision.

Kundel, H. L., & Nodine, C. F. (1978). Studies of eye movements and visual search in radiology. *Eye Movements and the Higher Psychological Functions*, 317–328.

Kundel, H. L., Nodine, C. F., Conant, E. F., & Weinstein, S. P. (2007). Holistic component of image perception in mammogram interpretation: Gaze-tracking study. *Radiology*, *242*(2), 396–402.

Lan, S., Ren, Z., Wu, Y., Davis, L. S., & Hua, G. (2020). In *SaccadeNet: A fast and accurate object detector* (pp. 10397–10406). Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Land, M. F. (1992). Predictable eye-head coordination during driving. *Nature*, *359*(6393), 318–320.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*(25), 3559–3565.

Land, M., & Tatler, B. (2009). *Looking and acting: Vision and eye movements in natural behaviour*. Oxford University Press.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). In *Microsoft coco: Common objects in context* (pp. 740–755). European conference on computer vision, Springer.

Lindsay, G. W., & Miller, K. D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. *ELife*, *7*, e38105.

Lindsay, G. W., Rubin, D. B., & Miller, K. D. (2019). *A simple circuit model of visual cortex explains neural and behavioral aspects of attention*. BioRxiv.

Liu, N., & Han, J. (2018). A deep spatial contextual long-term recurrent convolutional network for saliency detection. *IEEE Transactions on Image Processing*, *27*(7), 3264–3274.

Lotter, W., Kreiman, G., & Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. *ArXiv Preprint ArXiv*, *1605*08104.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.
Ma, W. J., & Peters, B. (2020). A neural network walks into a lab: Towards using deep nets as models for human behavior. *ArXiv Preprint ArXiv*, *2005*02181.
Malcolm, G. L., & Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes: Evidence from eye movements. *Journal of Vision*, *9*(11), 1–13.
Mall, S., Brennan, P. C., & Mello-Thoms, C. (2019). Can a Machine Learn from Radiologists' Visual Search Behaviour and Their Interpretation of Mammograms—A Deep-Learning Study. *Journal of Digital Imaging*, *32*(5), 746–760.
Maxfield, J. T., Stalder, W. D., & Zelinsky, G. J. (2014). Effects of target typicality on categorical search. *Journal of Vision*, *14*(12), 1–11.
Maxfield, J. T., & Zelinsky, G. J. (2012). Searching through the hierarchy: How level of target categorization affects visual search. *Visual Cognition*, *20*(10), 1153–1163.
Mirpour, K., Arcizet, F., Ong, W. S., & Bisley, J. W. (2009). Been there, seen that: A neural mechanism for performing efficient visual search. *Journal of Neurophysiology*, *102*(6), 3481–3491.
Müller, H. J., Heller, D., & Ziegler, J. (1995). Visual search for singleton feature targets within and across feature dimensions. *Perception & Psychophysics*, *57*(1), 1–17.
Munoz, D. P., & Everling, S. (2004). Look away: The anti-saccade task and the voluntary control of eye movement. *Nature Reviews Neuroscience*, *5*(3), 218–228.
Nako, R., Wu, R., & Eimer, M. (2014). Rapid guidance of visual search by object categories. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(1), 50.
Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, *45*(2), 205–231.
Norman, D. A., & Shallice, T. (1986). Attention to action. In *Consciousness and self-regulation* (pp. 1–18). Springer.
Olivers, C. N., Peters, J., Houtkamp, R., & Roelfsema, P. R. (2011). Different states in visual working memory: When it guides attention and when it does not. *Trends in Cognitive Sciences*, *15*(7), 327–334.
Papadopoulos, D. P., Clarke, A. D., Keller, F., & Ferrari, V. (2014). *Training object class detectors from eye tracking data*. Springer, pp. 361–376.
Papagiannopoulou, E. A., Chitty, K. M., Hermens, D. F., Hickie, I. B., & Lagopoulos, J. (2014). A systematic review and meta-analysis of eye-tracking studies in children with autism spectrum disorders. *Social Neuroscience*, *9*(6), 610–632.
Pomplun, M., Reingold, E. M., & Shen, J. (2003). Area activation: A computational model of saccadic selectivity in visual search. *Cognitive Science*, *27*(2), 299–312.
Posner, M. I. (1978). *Chronometric explorations of mind*. Lawrence Erlbaum.
Rao, R. P., Zelinsky, G. J., Hayhoe, M. M., & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research*, *42*(11), 1447–1463.
Reynolds, J. H., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron*, *26*(3), 703–714.
Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., et al. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770.
Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in Neural Information Processing Systems*, 3856–3866.
Schmidt, J., & Zelinsky, G. J. (2009). Search guidance is proportional to the categorical specificity of a target cue. *Quarterly Journal of Experimental Psychology*, *62*(10), 1904–1914.
Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition*, *80*(1–2), 1–46.
Serre, T., Wolf, L., & Poggio, T. (2005). *Object recognition with features inspired by visual cortex*. *2*. IEEE, pp. 994–1000.

Sheinberg, D. L., & Logothetis, N. K. (2001). Noticing familiar objects in real world scenes: The role of temporal cortical neurons in natural vision. *Journal of Neuroscience*, *21*(4), 1340–1350.

Shrivastava, A., & Gupta, A. (2016). In *Contextual priming and feedback for faster r-cnn* (pp. 330–348). *European conference on computer vision*, Springer.

Shrivastava, A., Sukthankar, R., Malik, J., & Gupta, A. (2016). Beyond skip connections: Top-down modulation for object detection. *ArXiv Preprint ArXiv*, *161*206851.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv*, *1409*, 1556.

Smith, T. J., & Henderson, J. M. (2011). Looking back at Waldo: Oculomotor inhibition of return does not prevent return fixations. *Journal of Vision*, *11*(1), 1–11.

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, *7*(1), 11–32.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766.

Tsotsos, J. K. (2011). *A computational perspective on visual attention*. MIT Press.

Tsotsos, J., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence*, *78*(1), 507–545.

Vickery, T. J., King, L.-W., & Jiang, Y. (2005). Setting up the target template in visual search. *Journal of Vision*, *5*(1), 81–92.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., et al. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), 860–868.

Wei, Z., Adeli, H., Zelinsky, G., Hoai, M., & Samaras, D. (2016). *Learned region sparsity and diversity also predict visual attention*.

Williams, D. E., Reingold, E. M., Moscovitch, M., & Behrmann, M. (1997). Patterns of eye movements during parallel and serial visual search tasks. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *51*(2), 151.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238.

Wolfe, J. (2020). In *Guided search 6.0: An upgrade with five forms of guidance, three types of functional visual fields, and two, distinct search templates*. Presentation at the 20th meeting of the vision sciences society.

Wolfe, J., Cain, M., Ehinger, K., & Drew, T. (2015). Guided Search 5.0: Meeting the challenge of hybrid search and multiple-target foraging. *Journal of Vision*, *15*(12), 1106.

Wolfe, J. M., & Gancarz, G. (1997). Guided search 3.0. In *Basic and clinical applications of vision science* (pp. 189–192). Springer.

Wolfe, J. M., Horowitz, T. S., Kenner, N., Hyle, M., & Vasan, N. (2004). How fast can you change your mind? The speed of top-down guidance in visual search. *Vision Research*, *44*(12), 1411–1426.

Wolfe, J. M., Horowitz, T. S., Palmer, E. M., Michod, K. O., & Van Wert, M. J. (2010). Getting into guided search. *Tutorials in Visual Cognition*, 93–119.

Wright, B. C. (2017). What Stroop tasks can tell us about selective attention from childhood to adulthood. *British Journal of Psychology*, *108*(3), 583–607.

Yamins, D., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.

Yang, Z., Huang, L., Chen, Y., Wei, Z., Ahn, S., Zelinsky, G., et al. (2020). In *Predicting goal-directed human attention using inverse reinforcement learning* (pp. 193–202). Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Yang, H., & Zelinsky, G. J. (2009). Visual search is guided to categorically-defined targets. *Vision Research*, *49*(16), 2095–2103.

Yarbus, A. L. (1967). Eye movements during perception of complex objects. In *Eye movements and vision* (pp. 171–211). Springer.

Yu, C.-P., Liu, H., Samaras, D., & Zelinsky, G. J. (2019). Modelling attention control using a convolutional neural network designed after the ventral visual pathway. *Visual Cognition*, 1–19.

Yu, Y., Mann, G. K., & Gosine, R. G. (2011). A goal-directed visual perception system using object-based top–down attention. *IEEE Transactions on Autonomous Mental Development*, *4*(1), 87–103.

Yu, C.-P., Maxfield, J. T., & Zelinsky, G. J. (2016). Searching for category-consistent features: A computational approach to understanding visual category representation. *Psychological Science*, *27*(6), 870–884.

Yu, C.-P., Samaras, D., & Zelinsky, G. J. (2014). Modeling visual clutter perception using proto-object segmentation. *Journal of Vision*, *14*(7), 1–16.

Zamir, A. R., Wu, T.-L., Sun, L., Shen, W. B., Shi, B. E., Malik, J., et al. (2017). In *Feedback networks* (pp. 1308–1317). Proceedings of the IEEE conference on computer vision and pattern recognition.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, *115*(4), 787.

Zelinsky, G. J. (2012). TAM: Explaining off-object fixations and central fixation tendencies as effects of population averaging during search. *Visual Cognition*, *20*(4–5), 515–545.

Zelinsky, G. J., Adeli, H., Peng, Y., & Samaras, D. (2013). Modelling eye movements in a categorical search task. *Phil Trans R Soc B*, 368(1628), 20130058.

Zelinsky, G. J., & Bisley, J. W. (2015). The what, where, and why of priority maps and their interactions with visual working memory. *Annals of the New York Academy of Sciences*, *1339*(1), 154–164.

Zelinsky, G. J., Chen, Y., Ahn, S., Adeli, H., Yang, Z., Huang, L., et al. (2020). Predicting Goal-directed Attention Control Using Inverse-Reinforcement Learning. *ArXiv Preprint ArXiv*, *2001*, 11921.

Zelinsky, G. J., Peng, Y., Berg, A. C., & Samaras, D. (2013). Modeling guidance and recognition in categorical search: Bridging human and computer object detection. *Journal of Vision*, *13*(3), 1–20.

Zelinsky, G. J., Peng, Y., & Samaras, D. (2013). Eye can read your mind: Decoding gaze fixations to reveal categorical search targets. *Journal of Vision*, *13*(14), 1–13.

Zelinsky, G. J., & Sheinberg, D. L. (1997). Eye movements during parallel–serial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(1), 244.

Zelinsky, G., Yang, Z., Huang, L., Chen, Y., Ahn, S., Wei, Z., et al. (2019). In *Benchmarking gaze prediction for categorical visual search* (pp. 828–836). 2019IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW). https://doi.org/10.1109/CVPRW.2019.00111.

Zhang, M., Feng, J., Ma, K. T., Lim, J. H., Zhao, Q., & Kreiman, G. (2018). Finding any Waldo with zero-shot invariant and efficient visual search. *Nature Communications*, *9*(1), 1–15.

Zhang, W., Yang, H., Samaras, D., & Zelinsky, G. J. (2006). A computational model of eye movements during object class detection. In Y. Weiss, B. Schölkopf, & J. C. Platt (Eds.), *Advances in neural information processing systems 18* (pp. 1609–1616). MIT Press.

Zheng, H., Fu, J., Mei, T., & Luo, J. (2017). In *Learning multi-attention convolutional neural network for fine-grained image recognition* (pp. 5209–5217). Proceedings of the IEEE international conference on computer vision.

Zoran, D., Chrzanowski, M., Huang, P.-S., Gowal, S., Mott, A., & Kohli, P. (2020). In *Towards robust image classification using sequential attention models* (pp. 9483–9492). Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.