Center bias outperforms image salience but not semantics in accounting for attention during scene viewing

Taylor R. Hayes[1] and John M. Henderson[1,2]
[1]Center for Mind and Brain, University of California, Davis
[2]Department of Psychology, University of California, Davis

How do we determine where to focus our attention in real-world scenes? Image saliency theory proposes that our attention is 'pulled' to scene regions that differ in low-level image features. However, models that formalize image saliency theory often contain significant scene-independent spatial biases. In the present studies, three different viewing tasks were used to evaluate whether image saliency models account for variance in scene fixation density based primarily on scene-dependent, low-level feature contrast, or on their scene-independent spatial biases. For comparison, fixation density was also compared to semantic feature maps (Meaning Maps; Henderson & Hayes, 2017) that were generated using human ratings of isolated scene patches. The squared correlations (R^2) between scene fixation density and each image saliency model's center bias, each full image saliency model, and meaning maps were computed. The results showed that in tasks that produced observer center bias, the image saliency models on average explained 23% less variance in scene fixation density than their center biases alone. In comparison, meaning maps explained on average 10% more variance than center bias alone. We conclude that image saliency theory generalizes poorly to real-world scenes.

Keywords: scene perception, center bias, saliency, semantics, meaning map

Real-world visual scenes are too complex to be taken in all at once (Tsotsos, 1991; Henderson, 2003). To cope with this complexity, our visual system uses a divide and conquer strategy by shifting our attention to different smaller subregions of the scene over time (Findlay & Gilchrist, 2003; Henderson & Hollingworth, 1999; Hayhoe & Ballard, 2005). This solution leads to a fundamental question in cognitive science: How do we determine where to focus our attention in complex, real-world scenes?

One of the most influential answers to this question has been visual salience. Image salience theory proposes that our attention is 'pulled' to visually salient locations that differ from their surrounding regions in semantically uninterpreted image features like color, orientation, and luminance (Itti & Koch, 2001). For example, a search array that contains a single red line among an array of green lines stands out and draws our attention (Treisman & Gelade, 1980; Wolfe, Cave, & Franzel, 1989; Wolfe, 1994). The idea of visual salience has been incorporated into many influential theories of attention (Wolfe & Horowitz, 2017; Itti & Koch, 2001; Wolfe et al., 1989; Treisman & Gelade, 1980) and formalized in various computational image saliency models (Itti, Koch, & Niebur, 1998; Harel, Koch, & Perona, 2006; Bruce & Tsotsos, 2009). These prominent image saliency models have influenced a wide range of fields including vision science, cognitive science, visual neuroscience, and computer vision (Henderson, 2007).

However, an often overlooked component of image saliency models is the role that image-independent spatial biases play in accounting for the distribution of scene fixations (Bruce, Wloka, Frosst, Rahman, & Tsotsos, 2015). Many of the most influential image saliency models exhibit significant image-independent spatial biases to account for observer center bias (Rahman & Bruce, 2015; Bruce et al., 2015). Observer center bias refers to the common empirical finding that human observers tend to concentrate their fixations more centrally when viewing scenes (Tatler, 2007). Tatler (2007) showed observer center bias is largely independent from scene content and viewing task, and suggested it may be the result of a basic orienting response, information processing strategy, or it may facilitate gist extraction for contextual guidance (Torralba, Oliva, Castelhano, & Henderson, 2006). Regardless of the source, these findings highlight the importance of taking observer center bias into account in evaluating models of scene attention.

This led us to ask a simple question: Are image saliency models actually predicting where we look in scenes based on low-level feature contrast, or are they mostly capturing that we tend to look more at the center than the periphery of scenes? The answer to this question is important because when image saliency models are successful in predicting fixation density, it is often implicitly assumed that scene-dependent, low-level feature contrast is responsible for this success in support of image guidance theory (Tatler, 2007; Bruce et al., 2015).

To answer this question, we compared how well three influential and widely cited image saliency models (Itti & Koch saliency model with Gaussian blur, Itti et al., 1998; Koch & Ullman, 1985; Harel et al., 2006; graph-based visual saliency model, Harel et al., 2006; and attention by information maximization saliency model, Bruce & Tsotsos, 2009) predicted scene fixation density relative to their respective image-independent center biases for three different scene viewing tasks: memorization, aesthetic judgment, and visual search. These image saliency models were chosen for two reasons. First, they are bottom-up image saliency models that allow us to cleanly dissociate low-level image features associated with image salience theory from high-level semantic features associated with cognitive guidance theory. Second, the chosen image saliency models each produce different degrees and patterns of spatial bias. The Itti and Koch and the graph-based visual saliency models both contain substantial image-independent spatial center biases with different profiles, while the attention by information maximization model is much less center-biased and served as a low-bias control. The memorization, aesthetic judgment, and visual search tasks were chosen because they produced varying degrees and patterns of observer center bias that allowed us to examine how the degree of observer center bias affects the performance of the various image saliency models.

Finally, as an additional analysis of interest, we compared each center bias baseline model and image saliency model to meaning maps (Henderson & Hayes, 2017, 2018). Meaning maps draw inspiration from two classic scene-viewing studies (Antes, 1974; Mackworth & Morandi, 1967). In these studies, images were divided into several regions and subjects were asked to rate each region based on how easy it would be to recognize (Antes, 1974) or how informative it was (Mackworth & Morandi, 1967). Critically, when a separate group of subjects freely viewed the same images, they mostly looked at the regions that were rated as highly recognizable or informative. Meaning maps scale up this general rating procedure using crowd-sourced ratings of thousands of isolated scene patches densely sampled at multiple spatial scales to capture the spatial distribution of semantic features, just as image saliency maps capture the spatial distribution of image features.

To summarize, the goal of the present article is to test whether image salience theory, formalized as image saliency models, offers a compelling answer to how we determine where to look in real-world scenes. We tested how well three different image saliency models accounted for fixation density relative to their respective center biases across three different tasks that produced varying degrees of observer center bias. The results showed that for tasks that produce observer center bias, image saliency models actually perform worse than their center bias alone. This finding suggests a serious disconnect between image salience theory and human attentional guidance in real-world scenes. In comparison, meaning maps were able to explain additional variance above and beyond center bias in all 3 tasks. These findings suggest image saliency models scale poorly to real-world scenes.

Method

Participants

The present study analyzes a corpus of data from 5 different groups of participants. Three different groups of students from the University of South Carolina (memorization, N=79) and the University of California, Davis (visual search, N=40; aesthetic judgment, N=53) participated in the eye tracking studies. Two different groups of Amazon Mechanical Turk workers (N=165) and University of California, Davis students (N=204) participated in the meaning map studies. All five studies were approved by the institutional review board at the university where they were collected. All participants in the eye tracking studies had normal or corrected to normal vision, were naïve concerning the purposes of each experiment, and provided written or verbal consent.

We have previously used the memorization study corpus to investigate individual differences in scan patterns in scene perception (Hayes & Henderson, 2017, 2018), as well as for an initial study of meaning maps (Henderson & Hayes, 2017, 2018). The observer center bias data and the comparisons to multiple image saliency and center bias models are presented here for the first time.

Stimuli

The study stimuli were digitized photographs of outdoor and indoor real-world scenes (See Figure 1a). The memorization study contained 40 scenes, the visual search study contained 80 scenes, and the aesthetic judgment study contained 50 scenes. In the visual search study 40 of the scenes contained randomly placed letter L targets (excluding the area within 2° of the pre-trial fixation cross) and 40 scenes contained no letter targets. Only the 40 scenes that did not contain letter targets were included in the analysis to avoid any contamination due to target fixations. The memorization and visual search study contained the same 40 scenes. The aesthetic judgment study shared 12 scenes with the memorization and visual search scene set.

Apparatus

Eye movements were recorded with an EyeLink 1000+ tower-mount eye tracker (spatial resolution 0.01°) sampling at 1000 Hz (SR Research, 2010b). Participants sat 85 cm away from a 21" monitor, so that scenes subtended approximately 27° x 20.4° of visual angle. Head movements were minimized using a chin and forehead rest. Although viewing was binocular, eye movements were recorded from the right eye. The experiment was controlled with SR Research Experiment Builder software (SR Research, 2010a).

Procedure

In the memorization study, subjects were instructed to memorize each scene in preparation for a later memory test, which was not administered. In the visual search study, subjects were instructed to search each scene for between 0 and 2 small embedded letter L targets and then respond with how many they found at the end of the trial. In the aesthetic judgment study, subjects were instructed to indicate how much they liked each scene on a 1-3 scale. For all three eye tracking studies, each trial began with fixation on a cross at the center of the display for 300 msec. Following central fixation, each scene was presented for 12 seconds while eye movements were recorded.

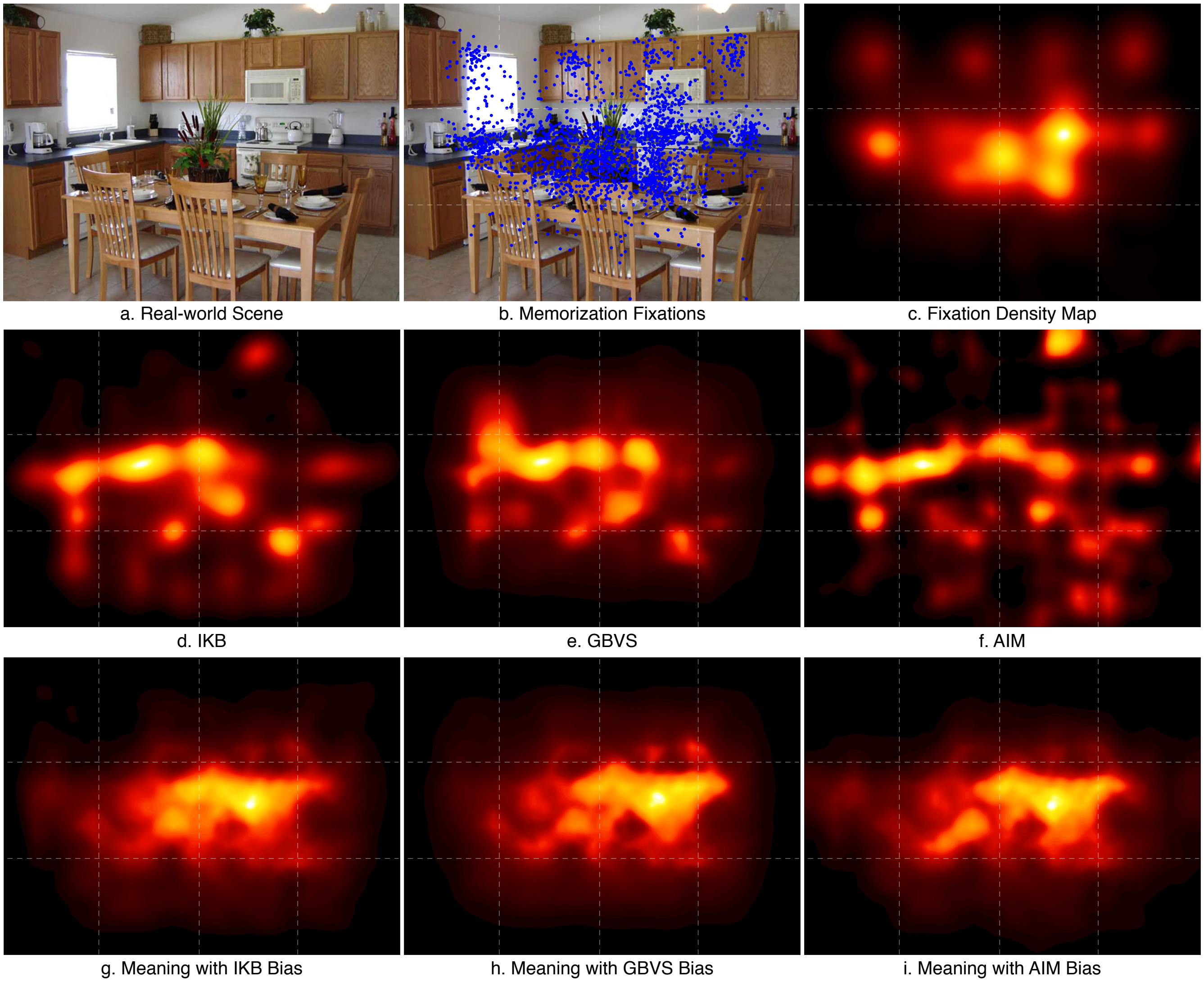

Figure 1. *A typical scene and the corresponding fixation density, image saliency, and meaning maps.* The top row shows a typical scene (a), the individual fixations produced by all participants in the memorization study (b), and the resulting fixation density map (c). The middle row shows the saliency maps produced by the Itti & Koch with blur saliency model (IKB, d), the Graph-based Visual saliency model (GBVS, e), and Attention by Information Maximization model (AIM, f). The bottom row shows the meaning maps with each of the corresponding image saliency model spatial biases applied (g, h, i). All maps were normalized using image histogram matching with the fixation density map (c) as the reference image. The dotted white lines are shown to make comparison across panels easier.

Eye Movement Data Processing

A 9-point calibration procedure was performed at the start of each session to map eye position to screen coordinates. Successful calibration required an average error of less than 0.49° and a maximum error of less than 0.99°. Fixations and saccades were segmented with EyeLink's standard algorithm using velocity and acceleration thresholds ($30/s$ and $9500°/s^2$).

Eye movement data were converted to text using the EyeLink EDF2ASC tool and then imported into MATLAB for analysis. Custom code was used to examine subject data for data loss from blinks or calibration loss based on mean percent signal across trials (Holmqvist, Nyström, Dewhurst, Jorodzka, & van de Weijer, 2015). In the memorization study, 14 subjects with less than 75% signal were removed, leaving 65 subjects for analysis that were tracked well, with an average signal of 91.7% (SD=5.5). In the aesthetic judgment study, 3 subjects with less than 75% signal were removed, leaving 50 subjects that were tracked well, with an average signal of 90.7% (SD=5.8). In the visual search study, 2 subjects with less than 75% signal were removed, leaving 38 subjects for analysis that were tracked well, with an average signal of 95.00% (SD=3.69). The first fixation in every trial was discarded as uninformative because it was constrained by the pretrial fixation cross.

Fixation Density Map

The distribution of scene attention was defined as the distribution of fixations within each scene. For each task, a fixation density map (Figure 1c) was generated for each scene across all subject fixations (Figure 1b). Following our previous work (Henderson & Hayes, 2017), the fixation frequency matrix for each scene was smoothed using a Gaussian low-pass filter with a circular boundary and a cutoff frequency of $-6dB$ to account for foveal acuity and eye-tracker error (Judd, Durand, & Torralba, 2012).

Image Saliency Maps

Saliency maps were generated for each scene using three different image saliency models. The Itti and Koch model with blur (IKB, Figure 1d) and the Graph-based Visual Saliency model (GBVS, Figure 1e) use local differences in image features including color, edge orientation, and intensity to compute a saliency map (Itti et al., 1998; Harel et al., 2006). The saliency maps for both the IKB and GBVS saliency models were generated using the Graph-based Visual Saliency toolbox with default GBVS settings and default IKB settings (Harel et al., 2006). The Attention by Information Maximization saliency model (AIM, Figure 1f) uses a different approach and computes an image saliency map based on each scene region's Shannon self-information (Bruce & Tsotsos, 2009). The AIM saliency maps were generated using the AIM toolbox with default settings and blur (Bruce & Tsotsos, 2009).

Meaning Maps

Meaning maps were generated as a representation of the spatial distribution of semantic information across scenes (Henderson & Hayes, 2017). Meaning maps were created for each scene by decomposing the scene into a dense array of overlapping circular patches at a fine spatial scale (300 patches with a diameter of 87 pixels) and coarse spatial scale (108 patches with a diameter of 207 pixels). Participants (*N*=369) provided ratings of thousands (31824) of scene patches based on how informative or recognizable they thought they were on a 6-point Likert scale. Patches were presented in random order and without scene context, so ratings were based on context-independent judgments. Each unique patch was rated by three unique raters.

A meaning map was generated for each scene by averaging the rating data at each spatial scale separately, then averaging the spatial scale maps together, and finally smoothing the average rating map with a Gaussian filter (i.e., Matlab 'imgaussfilt' with sigma=10).

Because meaning maps are generated based on context-independent random patch ratings, they by definition reflect content-dependent features. For comparison with the image saliency models, each image saliency model's spatial center bias was applied to the meaning maps for each scene by applying a pixel-wise multiplication with each image saliency model's center bias (See Figure 1g, h, i). This let us examine how the same center bias from each saliency model affected meaning map performance, allowing for a direct comparison of low-level image features and semantic features under the same center bias conditions.

Quantifying and Visualizing Center Bias

Each image saliency model has a unique scene-independent center bias. Therefore, the center bias was estimated for each image saliency model separately (i.e., IKB, GBVS, and AIM) using a large set of scenes from the MIT-1003 benchmark data set (Judd, Ehinger, Durand, & A., 2009). Specifically, we used the scene size that was most common in MIT-1003 data set (1024 x 768 px) and removed the 4 synthetic images resulting in 459 real-world scenes. For each image saliency model we generated a saliency map for each scene (459 scenes x 3 models = 1377 saliency maps). We then placed all the saliency maps on a common scale by normalizing each saliency map to have zero mean and unit variance.

In order to visualize each image saliency model's unique center bias we first computed the relative spatial bias across models (Bruce et al., 2015). That is, the relative spatial bias for each model was computed as the difference between the mean across all the saliency maps within each model (Figures 2a, 2b, 2c), minus the global mean across all the model saliency maps (Figures 2d). Figure 2e, 2f, and 2g show the resulting relative spatial biases for each image saliency model. This provides a direct visualization of how the different image saliency model biases compare relative to each other (Bruce et al., 2015).

We quantified the strength of the center bias in each image saliency model using the relative bias maps and a weight matrix that assigned weights according to center proximity. The relative saliency maps for each model (Figure 2e, 2f, and 2g) were jointly rescaled from 0 to 1 to maintain contrast changes. Next we needed to define how center bias was going to be weighted across image space. We computed the Euclidean distance from the center pixel to all other pixels, scaled it from 0 to 1, and then inverted it (Figure 2d). This served as a weight matrix representing center proximity. Each saliency model's bias was then simply the sum of the element-wise product of its relative bias map (Figure 2e, 2f, or 2g) and the center weight matrix (Figure 2h).

These same procedures were used to quantify observer center bias and visualize relative observer center bias for each eye tracking study (see Figure 3). The only difference is that the three studies took the place of the three saliency models, fixation density maps took the place of saliency maps, and the scenes that were viewed in each study were used instead of the MIT-1003 scenes. The smaller number of scenes in the eye tracking studies resulted in noisier estimates of the observer center bias maps (Figure 3) relative to the model center bias maps (Figure 2).

Map Normalization

The saliency and meaning maps were normalized using image-histogram matching in the same manner as Henderson and Hayes (2017, 2018). Histogram matching of the saliency

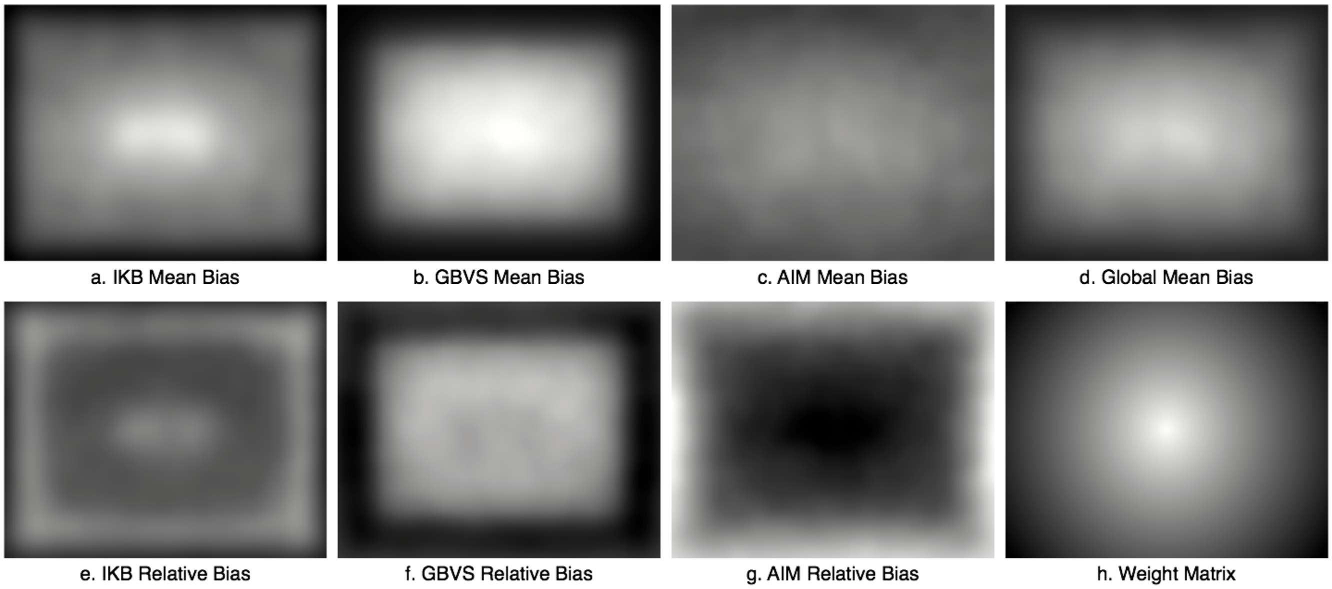

Figure 2. *Image saliency model mean bias, global bias, relative bias, and weight matrix.* The mean spatial bias is shown for each image saliency model, including (a) Itti and Koch with blur (IKB), (b) Graph-based Visual Saliency (GBVS), and (c) Attention by Information Maximization (AIM) for the MIT-1003 dataset. The relative spatial center bias for each model (e, f, g) shows how each saliency model differs relative to the global mean across all model saliency maps (d). Panel h shows the inverted Euclidean distance from the image center that served as the weight matrix for quantifying the degree of center bias in each model.

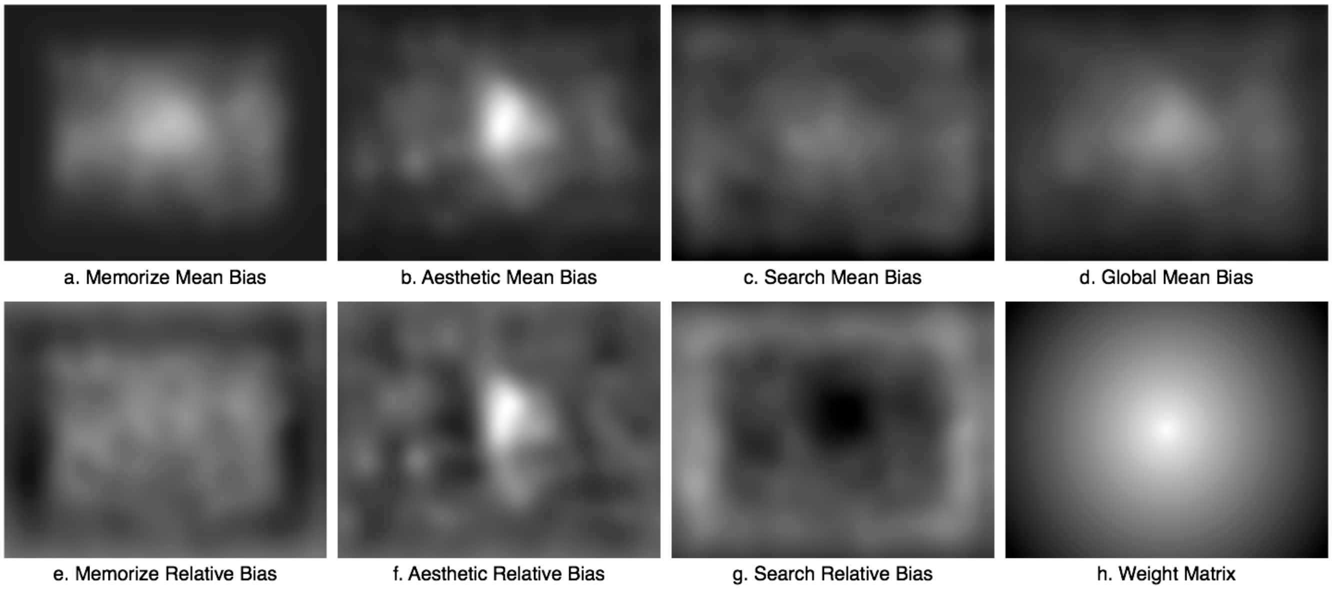

Figure 3. *Eye movement mean observer bias, global bias, relative bias, and weight matrix.* The mean observer bias is shown for each eye tracking study, including (a) scene memorization, (b) aesthetic judgment, and (c) visual search. The relative spatial center bias for each study (e, f, g) shows how each study differs relative to the global mean across all task fixation density maps (d). Panel h shows the inverted Euclidean distance from the image center that served as the weight matrix for quantifying the degree of observer center bias in each study.

and meaning maps was performed using the MATLAB function ‘imhistmatch’ from the Image Processing Toolbox. The fixation density map (Figure 1c) for each scene served as the reference image for the corresponding saliency (Figures 1d, 1e, 1f) and meaning (Figures 1g, 1h, 1i) maps.

Results

Image Saliency Model and Observer Spatial Biases

The image saliency models and the experimental tasks both produced varying degrees and patterns of spatial bias (Figure 2 and Figure 3). In Figures 2 and 3, panels a, b, and c show the average scene-independent spatial bias and panels e, f, and g show the relative spatial bias indicating how each model or task differed relative to all other models or tasks respectively. We quantified the strength of the center bias in each image saliency model and experimental task using the relative bias maps and a weight matrix (Figure 2h and Figure 3h) that assigned weights according to center proximity.

A comparison of the image saliency models showed clear differences in the degree and spatial profile of center bias in each model (Figure 2). GBVS displayed the strongest center bias followed by IKB (17.3% < GBVS) and AIM (47.4% < GBVS). The experimental tasks also produced varying degrees and amounts of observer center bias (Figure 3). The memorization task produced the most observer center bias followed by the aesthetic judgment (3% < memorization) and the visual search tasks (18% < memorization). It will be important to keep the relative strength of these spatial biases in mind as we examine model performance.

Model Performance

The main results are shown in Figure 4. For each study task (memorization, aesthetic judgment, and visual search), we computed the mean squared correlation (R^2) across all scene fixation density maps (circles) and each image saliency model (Figures 1d, 1e, 1f), each saliency model's center bias only (Figures 2a, 2b, 2c), and meaning maps with the same center bias as the image saliency model (Figures 1g, 1h, 1i). In this framework, the center bias only models serve as a baseline to measure how the addition of scene-dependent image saliency and scene-dependent semantic features affected performance. Two-tailed, paired sample t-tests were used to determine significance relative to the respective center bias only baseline models.

Figure 4a shows the memorization task results. We found that the three image saliency models each performed worse than their respective center biases alone. The full GBVS saliency model accounted for 8.1% less variance than the GBVS center bias model (t(39)=-3.14, $p < .01$, 95% CI $[-0.13, -0.03]$). The full IKB model accounted for 25.5% less variance than the IKB center bias model (t(39)=-9.13, $p < .001$, 95% CI $[-0.20, -0.31]$). Finally, the full AIM model accounted for 33.2% less variance than the AIM center bias model (t(39)=-12.02, $p < .001$, 95% CI $[-0.28, -0.39]$). It is worth noting that the full AIM model performed so poorly because its center bias is weaker than the GBVS and IKB models (recall Figure 2g). As a result, scene-dependent image salience played a much more prominent role in the AIM saliency maps to its detriment.

Figure 4b shows the aesthetic judgment task results. We found that again the three image saliency models each performed significantly worse than their respective center biases alone. The full GBVS saliency model accounted for 13.3% less variance than the GBVS center bias model (t(49)=-4.50, $p < .001$, 95% CI $[-0.07, -0.19]$). The full IKB model accounted for 25.8% less variance than the IKB center bias model (t(49)=-7.93, $p < .001$, 95% CI $[-0.19, -0.32]$). Finally, the full AIM model accounted for 32.5% less variance than the AIM center bias model (t(49)=-11.86, $p < .001$, 95% CI $[-0.27, -0.38]$).

Figure 4c shows the visual search task results. Recall that in the visual search task participants were searching for randomly placed letters, which greatly reduced the observer center bias (Figure 3g). We found that the three image saliency models each performed slightly better (GBVS, 4.8%; IKB, 6.8%; AIM, 5.3%) than their respective center biases alone in the visual search task (GBVS, t(39)=3.39, $p < .001$, 95% CI $[0.02, 0.7]$; IKB, t(39)=3.63, $p < .001$, 95% CI $[0.03, 0.11]$; AIM,t(39)=2.57, $p < .05$, 95% CI $[0.01, 0.09]$). This change is reflective of the much weaker observer center bias in the visual search task relative to the memorization and aesthetic judgment tasks. Together these factors greatly reduce the squared correlation of the center bias only model.

In contrast, the distribution of semantic features captured by meaning maps were always able to explain more variance than each center bias model alone. In the memorization task meaning maps explained on average 9.7% more variance than center bias alone (GBVS bias, t(39)=6.32, $p < .001$, 95% CI $[0.08, 0.15]$; IKB bias, t(39)=6.39, $p < .001$, 95% CI $[0.06, 0.12]$; AIM bias, t(39)=6.41, $p < .001$, 95% CI $[0.06, 0.12]$). In the aesthetic judgment task, meaning maps explained on average 10.3% more variance than center bias alone (GBVS bias, t(49)=3.60, $p < .001$, 95% CI $[0.04, 0.12]$; IKB bias, t(49)=6.65, $p < .001$, 95% CI $[0.08, 0.16]$; AIM bias, t(49)=5.93, $p < .001$, 95% CI $[0.07, 0.15]$). Finally, in the visual search task, meaning maps explained on average 10.0% more variance than the center bias only models (GBVS bias, t(39)=8.99, $p < .001$, 95% CI $[0.06, 0.10]$; IKB bias, t(39)=10.25, $p < .001$, 95% CI $[0.09, 0.14]$; AIM bias, t(39)=10.27, $p < .001$, 95% CI $[0.08, 0.13]$).

There has been some evidence suggesting that early attentional guidance may be more strongly driven by image salience than later attention (O'Connel & Walther, 2015; Anderson, Donk, & Meeter, 2016). Therefore, we performed a posthoc analysis to examine how the relationship between fixation density and each model varied as a function of viewing time. Specifically, we computed the correlation between the fixation density maps and each model in the same way as before, but instead of aggregating across all the fixations, we aggregated as a function of the fixations up to that point. That is, for each scene, we computed the fixation density map that contained only the first fixation for each subject, then the first

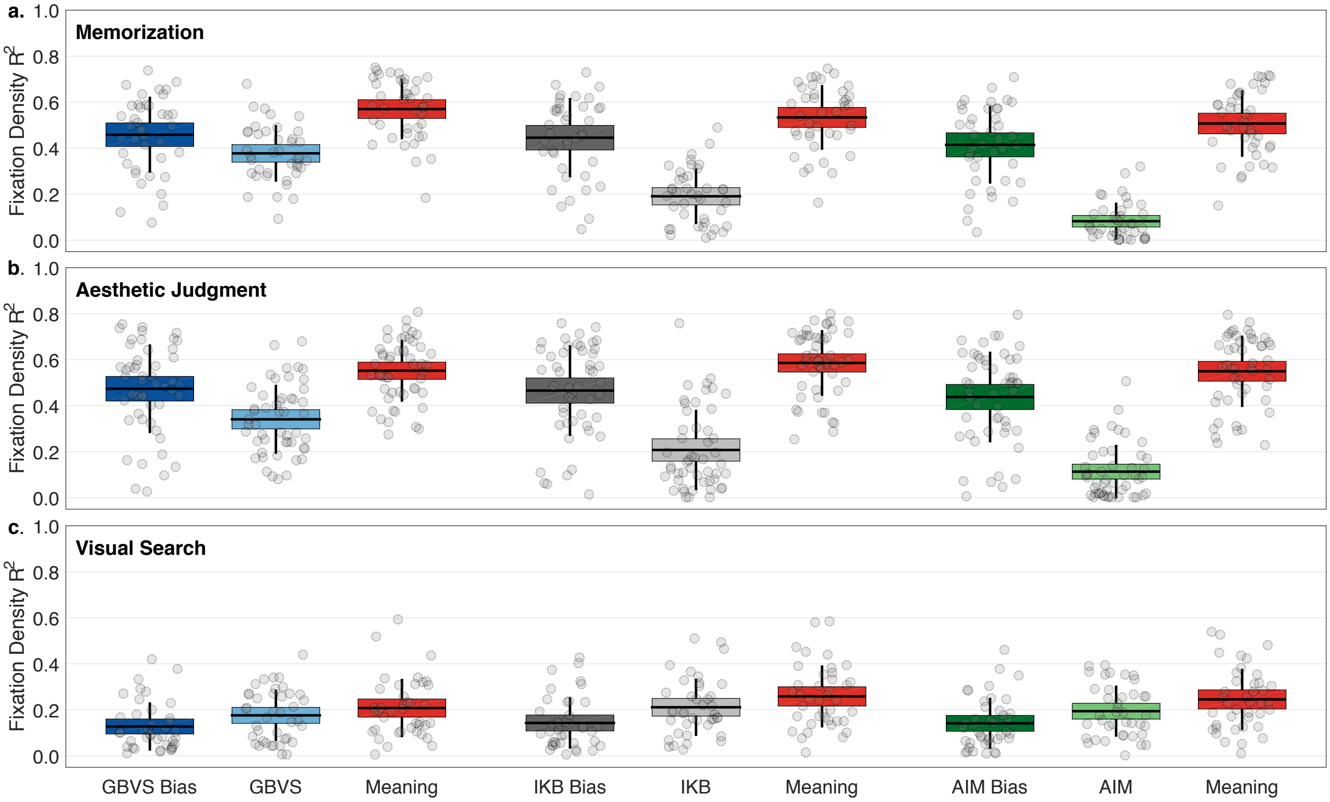

Figure 4. *Squared linear correlation between fixation density and maps across all scenes for each scene viewing task.* The scatter box plots show the squared correlation (R^2) between the scene fixation density maps and the saliency center bias maps (Graph-based visual salience, GBVS; Itti & Koch with blur, IKB; and Attention by Information Maximization saliency model, AIM), full saliency maps, and meaning maps for each scene task. The scatter box plots indicate the grand correlation mean (black horizontal line) across all scenes (circles), 95% confidence intervals (colored box) and 1 standard deviation (black vertical line).

and second fixation for each subject, and so on to generate the fixation density map for each scene at each time point. We then averaged the correlation values across scenes just as before.

Figure 5 shows how the squared correlation between fixation density and each model varied over time. The results are consistent with the main analysis shown in Figure 4. Specifically, Figure 5 shows that the respective center bias models are more strongly correlated with the first few fixations than the full image saliency models. Second, Figure 5 shows that in tasks that are center biased (i.e., memorization and aesthetic judgment) center bias performs better than the full image saliency models regardless of the viewing time, while in tasks that are less center biased (i.e., visual search) the image saliency gains a small advantage over center bias that accrues over time. Finally, Figure 5 shows that the meaning advantage over image salience observed in Figure 4 holds across the entire viewing period including even the earliest fixations. This finding is inconsistent with the idea that early scene attention is biased toward image saliency.

Taken together, our findings suggest that in tasks that produce observer center bias, adding low-level feature saliency actually explains less variance in scene fixation density than a simple center bias model, and that the same pattern holds for the earliest fixations ruling out an early saliency effect. This highlights that scene-independent center bias and not image salience is explaining most of the fixation density variance in these models. In comparison, meaning maps were consistently able to explain significant variance in fixation density above and beyond the center bias baseline models in each task.

Discussion

We have shown in a number of recent studies that image saliency is a relatively poor predictor of where people look in real-world scenes, and that it is actually scene semantics that guide attention (Henderson & Hayes, 2017, 2018; Henderson, Hayes, Rehrig, & Ferreira, 2018; Peacock, Hayes, & Henderson, 2019). Here we extend this research in a number of ways. First, the present work directly quantified the role of model center bias and observer center bias in overall performance of three image saliency models and meaning maps across three different viewing tasks. We found that for tasks that produced observer center bias (i.e., memorization and aesthetic judgment), the image saliency models actually performed significantly worse than their respective center biases alone, while the meaning maps performed

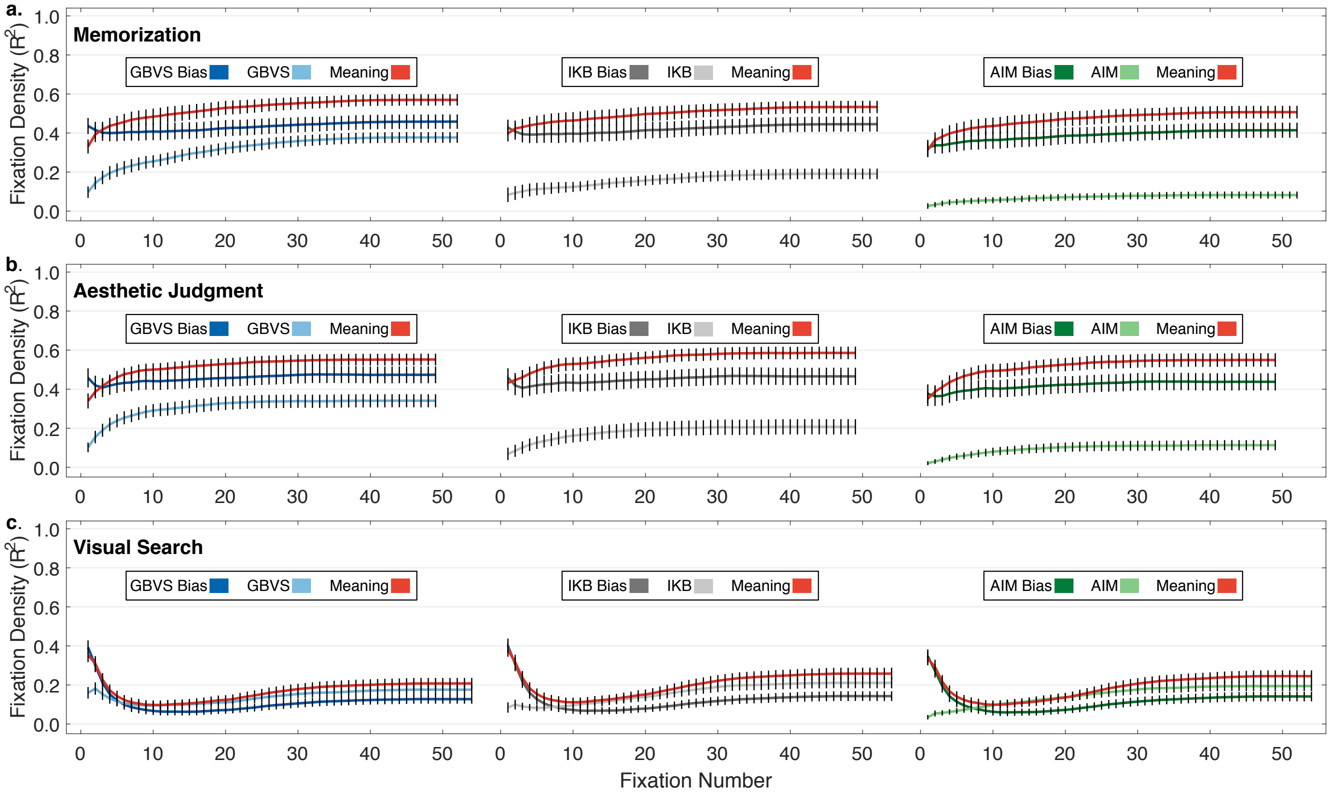

Figure 5. *Squared linear correlation between fixation density and maps across all scenes for each scene viewing task over time.* The line plots show the squared correlation (R^2) between the fixation density maps and the saliency center bias maps (Graph-based visual salience, GBVS; Itti & Koch with blur, IKB; and Attention by Information Maximization saliency model, AIM), full saliency maps, and meaning maps for each scene task over fixations. The lines indicate the grand mean across all scenes for each model up to each time point. The error bars indicate 95% confidence intervals.

significantly better than center bias alone regardless of task. Second, our previous work has exclusively used the graph-based saliency model (GBVS), whereas here we tested multiple image-based saliency models, demonstrating that center bias, image saliency, and meaning effects generalize across different models with different center biases. Finally, the temporal comparison of model center bias alone relative to the full saliency models shows clearly that early fixation density effects are predominantly center bias effects, not image saliency effects. Taken together these findings suggest that image salience theory does not offer a compelling account of where we look in real-world scenes.

So why does image salience theory instantiated as image saliency models struggle to account for variance beyond center bias? And why do semantic features succeed where salient image features fail? The most plausible explanation is the inherent difference between the semantically impoverished experimental stimuli that originally informed image saliency models, and semantically rich, real-world scenes.

The foundational studies that visual salience theory was built upon used singletons like lines and/or basic shapes that varied in low-level features like orientation, color, luminance, texture, shape, or motion (For review see Desimone & Duncan, 1995; Itti & Koch, 2000; Koch & Ullman, 1985). Critically, the singleton stimuli in these studies lacked any semantic content. The behavioral findings from these studies were then combined with new insights from visual neuroscience, such as center-surround receptive field mechanisms (Allman, Miezin, & McGuinness, 1985; Desimone, Schein, Moran, & Ungerleider, 1985; Knierim & Essen, 1992) and inhibition of return (Klein, 2000), to form the theoretical and computational basis for image saliency modeling (Koch & Ullman, 1985). When image saliency models were then subsequently applied to real-world scene images and found to account for a significant amount of variance in scene fixation density, it was taken as evidence that the visual salience theory scaled to complex, real-world scenes (Borji, Parks, & Itti, 2014; Borji, Sihite, & Itti, 2013; Harel et al., 2006; Parkhurst, Law, & Niebur, 2002; Itti & Koch, 2001; Koch & Ullman, 1985; Itti et al., 1998). The end result is that visual salience became a dominant theoretical paradigm for understanding attentional guidance not just in simple search arrays but in complex, real-world scenes (Henderson, 2007).

Our findings add to a growing body of evidence that attention in real-world scenes is not guided primarily by image salience, but rather by scene semantics. First, our results add to converging evidence that a number of widely used image saliency models account for scene attention primarily through their scene-independent spatial biases, rather than low-level feature contrast during free viewing (Kümmerer, Wallis, & Bethge, 2015; Bruce et al., 2015). Our findings generalize this effect to three additional scene viewing tasks: memorization, aesthetic judgment, and visual search tasks. Second, our results show that meaning maps are capable of explaining additional variance in overt attention beyond center bias in all three tasks. These results add to a number of recent studies that indicate that scene semantics are the primary factor guiding attention in real-world scenes (Henderson & Hayes, 2017, 2018; Henderson et al., 2018; Peacock et al., 2019; de Haas, Iakovidis, Schwarzkopf, & Gegenfurtner, 2019).

In terms of practical implications, our results together with previous findings (Tatler, 2007; Bruce et al., 2015; Kümmerer et al., 2015) suggest that image saliency model results should be interpreted with caution when used with real-world scenes as opposed to singleton arrays or other simple visual stimuli. Therefore, moving forward, it is prudent when using image saliency models with scenes to determine the degree of center bias in the fixation data and quantify the role center bias is playing in the image saliency model performance. These quantities can be measured and visualized using the aggregate map-level methods used here or other recently proposed methods (Kümmerer et al., 2015; Nuthmann, Einhäuser, & Schütz, 2017; Bruce et al., 2015). This will allow researchers to determine the relative contribution of scene-independent spatial bias and scene-dependent image salience when interpreting their data.

So where does this leave us? While image saliency theory and models offer an elegant framework based on biologically-inspired mechanisms, much of the behavioral work suggesting that low-level image feature contrast guides overt attention relies heavily on the use of semantically impoverished visual stimuli. Our results suggest that image saliency theory and models do not scale well to complex, real-world scenes. Indeed, we found prominent image saliency models actually did significantly worse than their center biases alone in multiple studies. This suggests something critical is missing from image saliency theory and models of attention when they are applied to real-world scenes. Our previous and current results suggest that what is missing are scene semantics.

Open Practices Statement

Data are available from the authors upon reasonable request. None of the studies were preregistered.

References

Allman, J., Miezin, F. M., & McGuinness, E. (1985). Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annual review of neuroscience*, *8*, 407-30.

Anderson, N. C., Donk, M., & Meeter, M. (2016). The influence of a scene preview on eye movement behavior in natural scenes. *Pscyhonomic Bulletin & Review*, *23*(6), 1794–1801.

Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology*, *103*(1), 62-70.

Borji, A., Parks, D., & Itti, L. (2014). Complementary effects of gaze direction and early saliency in guiding fixations during free viewing. *Journal of Vision*, *14*(13), 1-32.

Borji, A., Sihite, D. N., & Itti, L. (2013). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, *22*(1), 55–69.

Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, *9*(3), 1-24.

Bruce, N. D., Wloka, C., Frosst, N., Rahman, S., & Tsotsos, J. K. (2015). On computational modeling of visual saliency: Examining what's right and what's left. *Vision Reearch*, *116*, 95–112.

de Haas, B., Iakovidis, A. L., Schwarzkopf, D. S., & Gegenfurtner, K. R. (2019). Individual differences in visual salience vary along semantic dimensions. *Proceedings of the National Academy of Sciences*, *116*(24), 11687–11692. Retrieved from `https://www.pnas.org/content/116/24/11687` doi: 10.1073/pnas.1820553116

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, *18*, 193-222.

Desimone, R., Schein, S. J., Moran, J. P., & Ungerleider, L. G. (1985). Contour, color and shape analysis beyond the striate cortex. *Vision Research*, *25*, 441-452.

Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing*. Oxford: Oxford University Press.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based Visual Saliency. In *Neural information processing systems* (pp. 1–8).

Hayes, T. R., & Henderson, J. M. (2017). Scan patterns during real-world scene viewing predict individual differences in cognitive capacity. *Journal of Vision*, *17*(5), 1-17.

Hayes, T. R., & Henderson, J. M. (2018). Scan patterns during scene viewing predict individual differences in clinical traits in a normative sample. *PLoS ONE*, *13*(5), 1–16.

Hayhoe, M. M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498–504.

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science*, *16*, 219–222.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes rereveal by meaning maps. *Nature Human Behaviour*, *1*, 743–747.

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision*, *18*(6:10), 1-18.

Henderson, J. M., Hayes, T. R., Rehrig, G., & Ferreira, F. (2018). Meaning guides attention during real-world scene description. *Scientific Reports*, *8*, 1–9.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, *50*, 243–271.

Holmqvist, K., Nyström, R., M.and Andersson, Dewhurst, R., Jorodzka, H., & van de Weijer, J. (2015). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489–1506.

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

Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. In *Mit technical report.*

Judd, T., Ehinger, K. A., Durand, F., & A., T. (2009). Learning to predict where humans look. *2009 IEEE 12th International Conference on Computer Vision*, 2106-2113.

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences*, *4*, 138–147.

Knierim, J. J., & Essen, D. C. V. (1992). Neuronal responses to static texture patterns in area v1 of the alert macaque monkey. *Journal of neurophysiology*, *67 4*, 961-80.

Koch, C., & Ullman, U. (1985). Shifts in selective visual attention: Towards a underlying neural circuitry. *Human Neurobiology*, *4*, 219-227.

Kümmerer, M., Wallis, T. S., & Bethge, M. (2015). Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences of the United States of America*, *112 52*, 16054-9.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics*, *2*(11), 547–552.

Nuthmann, A., Einhäuser, W., & Schütz, I. (2017). How well can saliency models predict fixation selection in scenes beyond central bias? a new approach to model evaluation using generalized linear mixed models. In *Front. hum. neurosci.*

O'Connel, T. P., & Walther, D. B. (2015). Dissociation of salience-driven and content-driven spatial attention to scene category with predictive decoding of gaze patterns. *Journal of Vision*, *15*(5), 1–13.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*, 102-123.

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019). Meaning guides attention during scene viewing even when it is irrelevant. *Attention, Perception, & Psychophysics*, *81*, 20-34.

Rahman, S., & Bruce, N. (2015). Visual saliency prediction and evaluation across different perceptual tasks. *PLOS ONE*, *10*(9), e0138053.

SR Research. (2010a). *Experiment Builder user's manual*. Mississauga, ON: SR Research Ltd.

SR Research. (2010b). *EyeLink 1000 user's manual, version 1.5.2*. Mississauga, ON: SR Research Ltd.

Tatler, B. W. (2007). The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14), 1–17.

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*, 766–786.

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, *12*, 97–136.

Tsotsos, J. K. (1991). Is complexity theory appropraite for analysing biological systems? *Behavioral and Brain Sciences*, *14*(4), 770-773.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, *1 2*, 202-38.

Wolfe, J. M., Cave, K. R., & Franzel, S. (1989). Guided search: an alternative to the feature integration model for visual search. *Journal of experimental psychology. Human perception and performance*, *15 3*, 419-33.

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, *1*, 1–8.