Version of Record: https://www.sciencedirect.com/science/article/pii/S0010027722002190
Manuscript_1215dd84609b1614299ef7e9be02dbf9

Scene inversion reveals distinct patterns of attention to semantically interpreted and uninterpreted features

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

Abstract

Semantic guidance theories propose that attention in real-world scenes is strongly associated with semantically informative scene regions. That is, we look where there are recognizable and informative objects that help us make sense of our visual environment. In contrast, image guidance theories propose that local differences in semantically uninterpreted image features such as luminance, color, and edge orientation primarily determine where we look in scenes. While it is clear that both semantic guidance and image guidance play a role in where we look in scenes, the degree of their relative contributions and how they interact with each other remains poorly understood. In the present study, we presented real-world scenes in upright and inverted orientations and used general linear mixed effects models to understand how semantic guidance, image guidance, and observer center bias were associated with fixation location and fixation duration. We observed distinct patterns of change under inversion. Semantic guidance was severely disrupted by scene inversion, while image guidance was mildly impaired and observer center bias was enhanced. In addition, we found that fixation durations for semantically rich regions decreased when viewing inverted scenes relative to upright scene viewing, while fixation durations for image salience and center bias were unaffected by inversion. Together these results provide important new constraints on theories and computational models of attention in real-world scenes.

Keywords: scene perception, semantic guidance, image guidance, inversion, eye movements

1. Introduction

We process our complex visual world by shifting our overt attention to prioritize some scene regions over others (Hayhoe & Ballard, 2005; Henderson, 2003, 2011). However, how we determine which scene regions to prioritize for attention remains a fundamental question in cognitive science. Image guidance theories propose that attention is primarily guided by local contrasts in semantically uninterpreted image features such as luminance, color, and edge orientation (Itti & Koch, 2001; Koch & Ullman, 1985; Parkhurst, Law, & Niebur, 2002). In contrast, semantic guidance theories propose that attention is primarily guided by scene semantics, where attention is guided by the cognitive system to scene regions that are recognizable, informative, and relevant to our current goals (Henderson, 2007; Henderson, Brockmole, Castelhano, & Mack, 2007; Henderson, Malcolm, & Schandl, 2009; Henderson,

2011). Therefore, the key difference between image-guidance and semantic-guidance theories is the degree to which semantically uninterpreted low-level image features and semantic representations guide attention in scenes.

Much of the research on how attention is guided in scenes centers on or is influenced by image-guidance theory in part because it is computationally tractable. Computational image saliency models use local image feature contrasts (e.g., luminance, color, and edge orientation) that are computed at multiple spatial scales and pooled to form an image saliency map (Borji, Parks, & Itti, 2014; Harel, Koch, & Perona, 2006; Itti & Koch, 2001; Parkhurst et al., 2002). The image saliency map provides a distribution of image salience for every pixel in the scene image and reflects the predicted distribution of attention for that scene. Critically, computing an image saliency map requires no semantic knowledge of the scene category or the objects within it. In comparison, semantic-guidance theory proposes that semantic knowledge of the scene category, the objects it contains, and/or the goals of the viewer are the primary determinants of the attentional priority in scenes (Buswell, 1935; Hayhoe & Ballard, 2005; Henderson & Hollingworth, 1999; Henderson, 2003). However, unlike image-guidance theory, there are no computational models that can generate a semantic analogue of an image saliency map from only a scene image due to the inherent difficulty in modeling the complexities of human semantic knowledge. For our purposes, semantic-guidance and image-guidance are very constrained terms, and simply distinguish between attention guided by scene features that are semantically interpreted versus those that are based on semantically uninterpreted low-level image properties[1].

A large and growing body of evidence supports the semantic guidance of scene attention. This includes research demonstrating that a viewer's task (Yarbus, 1967; Tatler, Hayhoe, Land, & Ballard, 2011; Rothkopf, Ballard, & Hayhoe, 2007; Einhäuser, Rutishauser, & Koch, 2008; Castelhano, Mack, & Henderson, 2009) and scene semantics (Potter, 1975; Biederman, 1972; Võ, Boettcher, & Draschkow, 2019; Williams & Castelhano, 2019; Wu, Wick, & Pomplun, 2014; Hwang, Wang, & Pomplun, 2011; Malcolm, Groen, & Baker, 2016; de Haas, Iakovidis, Schwarzkopf, & Gegenfurtner, 2019; Henderson & Hayes, 2017; Hayes & Henderson, 2021b) are the primary determinants of attention in scenes. Because there is no computational model of scene semantics (though recent advancements are being made in that direction Hayes & Henderson, 2021b), the association between scene semantics and attention has typically been studied by actively manipulating a small number of isolated objects in each scene (Loftus & Mackworth, 1978; Biederman, Mezzanotte, & Rabinowitz, 1982; Boyce, Pollatsek, & Rayner, 1989; Hollingworth & Henderson, 1998; Henderson, Weeks, & Hollingworth, 1999; Castelhano & Heaven, 2011; Võ & Henderson, 2011). In these paradigms, isolated scene objects are manipulated to be either more or less semantically consistent with the broader scene category. While these previous studies provide important direct demonstrations of the effect of scene-object semantics on attention, they are all spatially limited to only small portions of the broader scene.

A more complete understanding of the role semantic features on scene attention requires a complete map of the distribution of semantic features across entire scenes (i.e., a semantic map). Therefore, to address this spatial limitation in studying scene semantics, we introduced a technique called 'meaning mapping' that uses human raters to build a map of different local semantic features across entire scenes (Henderson & Hayes, 2017, 2018; Rehrig, Peacock, Hayes, Henderson, & Ferreira, 2020). The meaning mapping idea is simple: use human raters' rich semantic knowledge to tell us how different semantic features are distributed across scenes. In this way, a meaning map serves as a semantic analogue of an image saliency map and allows us to examine how semantically interpreted features are associated with attention across the entire scene (Henderson & Hayes, 2017).

The meaning mapping approach takes a scene and splits it into small circular patches at multiple spatial scales and then uses crowd-sourced ratings of these patches to estimate how informative and recognizable each scene patch is in isolation (Henderson & Hayes, 2017). These isolated patch ratings

[1]Image-guidance and semantic-guidance theory should not be conflated with 'bottom-up' and 'top-down' processing which have much broader and varied connotations in the attention literature.

are then combined back into their location in the scene to form a 'meaning map' that provides an estimate of the local semantic density at every location in the scene. Scene patches that are rich in recognizable semantic information are rated as highly meaningful (e.g., a cluttered counter top in a kitchen), while scene regions that are unrecognizable and/or contain very little information (e.g., a patch of texture or a white wall) are rated as very low meaning. Between these two extremes exists a rich continuum of patches of varying degrees of meaning (for an example of ratings produced by a typical human rater see https://osf.io/yt2dk/). Meaning maps have been used as a tool to demonstrate that across entire scenes local semantic density is one of the strongest predictors of attention in a wide range of scene tasks including scene memorization (Henderson & Hayes, 2017, 2018), visual search (Hayes & Henderson, 2019), free viewing (Peacock, Hayes, & Henderson, 2019b), scene description (Henderson, Hayes, Rehrig, & Ferreira, 2018), brightness estimation (Peacock, Hayes, & Henderson, 2019a), and even when the scene semantics are task-independent (Hayes & Henderson, 2019). While this work highlights the importance of generating a full map of semantic features, one limitation of this body of work is that it is largely correlational unlike the previous work that actively manipulates scene-object consistency.

A complementary approach to studying the role of semantic guidance and image guidance across entire scenes is to manipulate the entire scene in ways which should impact each of them differently. Previous work suggests that inverting scenes and/or objects makes a scene harder to identify, scene changes harder to detect, and object properties more difficult to extract (Shore & Klein, 2000; Rock, 1974; Kelley, Chun, & Chua, 2003; Epstein, Higgins, Parker, Aquirre, & Cooperman, 2006; Peterson & Gibson, 1994; Jolicoeur, 1988; Rock & DiVita, 1987; Tarr & Pinker, 1990). Importantly, semantically uninterpreted image feature contrasts remain the same when a scene is inverted. Therefore, scene inversion has a number of appealing qualities for studying the distinct roles of semantic guidance and image guidance in scenes. First, inversion is an active manipulation that should disrupt semantic guidance while leaving image guidance intact. Second, scene inversion provides a strong control for low-level image features since the image feature contrasts are identical in the upright and inverted scene viewing conditions. Finally, scene inversion manipulates the entire scene, and so, in conjunction with image saliency maps and meaning maps, it will allow us to estimate the degree of change across an entire scene for semantically uninterpreted low-level image features and semantically interpreted features for the first time.

In the present study, we actively manipulated scenes using a scene inversion paradigm and measured the effect on semantic-guidance, image-guidance, and observer center bias using a mixed-effects modeling approach. There are two important questions we wish to answer with this approach. First, to what degree are local scene semantics impaired by inversion across entire scenes? Previous studies have shown decrements in manipulated local regions, but these studies do not provide an estimate of how attention to local semantic features is affected globally across entire scenes. Second, if scene semantics are significantly disrupted by scene inversion, does this disruption then modulate image guidance and/or observer center bias? Since computational models of scene attention produce full maps of prediction, it is important to know how attention to both image features and critically semantic features are globally affected by inversion and how they interact when one is disrupted. Answering these questions has the potential to provide important new constraints on both theory and computational models of scene attention. That is, in addition to how attention operates in upright viewing, theories of scene attention and computational models will also have to be able to account for how attention to local semantic density, image salience, and observer center bias change with scene inversion.

2. Method

2.1. Eye tracking study

2.1.1. Participants. University of California, Davis undergraduate students (age mean=20.1, standard deviation=1.7) with normal or corrected-to-normal vision participated in the eye tracking (N=40) study in exchange for course credit. All participants were naive concerning the purposes of the experiment and provided verbal or written informed consent as approved by the University of California, Davis Institutional Review Board.

2.1.2. Stimuli. Each participant in the eye tracking study viewed 102 real-world scene images in upright and inverted orientations. The 102 scenes consisted of a mix of indoor and outdoor scenes.

2.1.3. Apparatus. Participant eye movements were recorded using 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 and viewed scenes that subtended approximately 27° x 20° 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 display presentation was controlled with SR Research Experiment Builder software (SR Research, 2010a).

2.1.4. Procedure. Each participant viewed 102 scenes while performing a scene memorization task. Participants were instructed to memorize each scene for a later memory test, but no memory test was administered. These task instructions were used to provide a concrete viewing task to keep participants consistently engaged throughout the experiment. Each trial began with a fixation on a cross at the center of the display for 300 ms followed by a scene presented for 6 seconds. The main manipulation was that each participant viewed half the scenes upright and half the scenes inverted, counterbalanced across participants. The upright and inverted scenes were presented in a different random order for each participant to control for presentation order and expectancy effects.

2.1.5. Eye tracking calibration and data quality. 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 $9500°/s^2$). A drift correction was performed before each trial and recalibrations were performed as needed. The recorded eye tracking data were examined for data artifacts from excessive blinking or calibration loss by calculating the mean percent signal across trials (Holmqvist, Nyström, Dewhurst, Jorodzka, & van de Weijer, 2015). Five subjects with less than 75% signal were removed, leaving 35 subjects that were tracked well (signal mean=91.9%, SD=4.7%).

The remaining participants (N=35) produced an eye-movement data set that contained 61260 fixations with an average of 1750 fixations per participant. The average participant fixation duration was 267 msec (SD=155 msec).

2.2. Meaning map rating study

2.2.1. Participants. University of California, Davis undergraduate students (N=416; age mean=20.2, standard deviation=1.7) with normal or corrected-to-normal vision participated in the meaning rating study for course credit. All participants were naive concerning the purposes of the experiment and provided verbal or written informed consent as approved by the University of California, Davis Institutional Review Board.

2.2.2. Stimuli. Each participant in the meaning rating study viewed and rated 300 random isolated, small circular scene regions taken from the same set of 102 scenes from the eye tracking study.

2.2.3. Procedure. Meaning maps were generated for each scene as a representation of the spatial distribution of local semantic density (Henderson & Hayes, 2017, 2018; see https://osf.io/654uh/ for code and complete rater instructions, and https://osf.io/ptsvm/ for the 102 scene meaning maps). A meaning map was created for each scene by cutting 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). Raters (N=416) viewed 300 isolated scene patches and provided ratings using a 6-point Likert scale based on how informative or recognizable they thought the content of each patch was (Henderson & Hayes, 2017; Mackworth & Morandi, 1967). Patches were presented in random order and without scene context, so ratings were based on context-independent judgments. Each unique patch was rated by 3 unique raters. A meaning map (Figure 1b) was generated for each scene by averaging the rating data at each spatial scale separately, then averaging the spatial scale maps together, and then smoothing the grand average rating map with a Gaussian filter (i.e., Matlab 'imgaussfilt' with $\sigma = 10$, FWHM=23 pixels).

2.2.4. Meaning ratings for inverted scenes. To assess whether meaning map ratings were significantly different for inverted scenes compared to upright scenes, we randomly sampled approximately one third of the scenes (33 scenes) and meaning mapped them using the inverted scenes (N=135). We then directly compared the inverted patch meaning ratings to the upright patch meaning ratings. The results showed no significant difference ($t(13463)$=0.213, p=0.83, 95% CI [-0.006 0.007]) in meaning ratings across the 13464 upright and inverted scene patches that were rated in these 33 scenes (i.e., 300 fine patches and 108 coarse patches per scene). The correlation across all the upright and inverted patch ratings was also high (R=0.908). Since there was not a significant difference between upright and inverted meaning patch ratings, the upright meaning map for each scene (e.g., Fig. 1b) was simply inverted (Fig. 1f) to serve as the map of local semantic density for the inverted viewing condition.

2.3. Additional feature maps

2.3.1 Image Saliency Map. An image saliency map was also generated for each scene (Figure 1c) using the Graph-based Visual Saliency (GBVS) toolbox with default settings (Harel et al., 2006). We chose the GBVS model because it is based on known low-level mechanisms of the human visual system (Borji & Itti, 2013; Itti, Koch, & Niebur, 1998; Itti & Koch, 2001) and is one of the best performers among low-level image saliency models (Walther & Koch, 2006). In comparison, state-of-the-art deep neural network models learn where people attend in scenes from training on scene fixation data over object features and are known to contain a mix of low-level, mid-level, and high-level features (Hayes & Henderson, 2021a; Henderson, Hayes, Peacock, & Rehrig, 2021). Therefore, a pure low-level image saliency model like GBVS provides a better estimate of how semantically uninterpreted image features are affected by scene inversion than a deep saliency model.

Since local contrasts in luminance, color, and orientation are unaffected by scene inversion, the image saliency map (Fig. 1c) was simply inverted (Fig. 1g) to serve as the image salience map for the inverted viewing condition.

2.3.2. Center Proximity Map. A center proximity map served as a global representation of how far each location in the scene was from the scene center. Specifically, the center proximity map measured the inverted Euclidean distance from the center pixel of the scene to all other pixels in the scene image (Figure 1d, h). The center proximity map (Hayes & Henderson, 2021b) was used to explicitly control for the general bias for observers to look more centrally than peripherally in scenes, independent of the underlying scene content (Tatler, 2007; Hayes & Henderson, 2020) and was therefore identical in the upright (Fig. 1d) and inverted (Fig. 1h) viewing conditions.

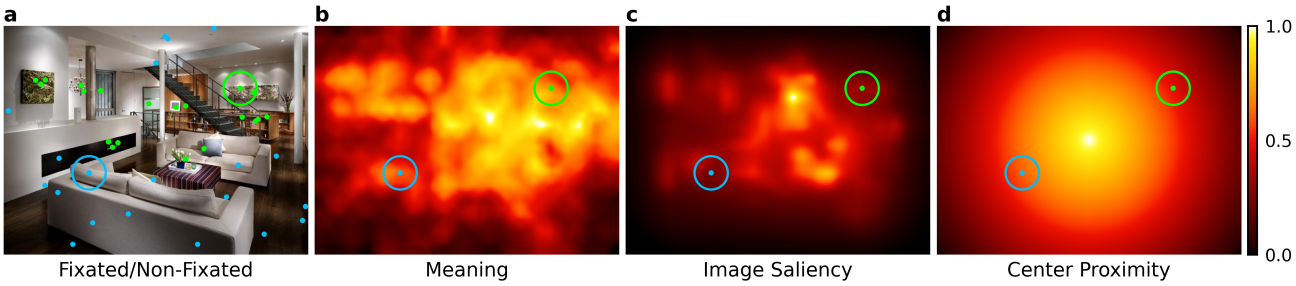

Figure 1. Scene viewing conditions, eye movement data, and feature maps. Each scene was presented in a upright (a) and inverted (e) orientation counterbalanced across viewers. The green dots show the fixation locations for a typical viewer and the cyan dots indicate randomly sampled non-fixated regions that indicate where each subject did and did not look (a, e). Together these locations provide an account of the regions in each scene that did and did not capture each subject's attention in the upright and inverted viewing conditions. Each fixated and non-fixated location was used to compute a mean value for each feature map, both for the upright condition (b, c, d) and the inverted condition (f, g, h), across a 3° window (shown as circles around an example fixated/non-fixated location).

2.4. General linear mixed effects models (GLMM) of eye movement behavior

2.4.1. Fixated and non-fixated scene locations. We modeled the association between the eye movement data and the different feature maps by comparing where each subject looked in each scene to where they did not look (Hayes & Henderson, 2021b; Nuthmann, Einhäuser, & Schütz, 2017). Specifically, for each region a subject fixated, we computed the mean value for each feature map for each viewing condition (upright: Figures 1b, 1c, 1d; inverted: Figures 1f, 1g, 1h) by taking the average over a 3° window around each fixation (Figure 1, neon green locations). To represent the scene regions that were not associated with overt attention, for each individual subject, we randomly sampled an equal number of scene locations where that subject did not look in each scene they viewed (Figure 1, cyan locations). The only constraint for the random sampling of the non-fixated scene regions was they could not overlap with any of the fixated 3° windows, which reflects a logical constraint that for a given scene viewed by a given subject, no scene region can be both fixated and not-fixated. This sampling procedure was performed separately for each individual scene viewed by each individual subject.

2.4.2. Fixation location general linear mixed effects models. We then used a general linear mixed-effects model (GLMM) approach to estimate the association between the fixated and non-fixated scene regions, our feature maps, and viewing condition (upright vs. inverted) using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) in R (R Core Team, 2017). All continuous predictors (i.e., meaning, GBVS, and center proximity) were standardized to have mean 0 and standard deviation of 1 prior to model fitting and the *glmer* function with a binomial distribution, logit link function, and the default optimizer (bobyqa and Nelder Mead) were used for fitting. A mixed effects modeling approach has a couple of important advantages. First, it does not require aggregating the eye movement data at the subject or scene-level like ANOVA or map-level correlations; instead, both subject and scene can be explicitly modeled as random effects. Second, the GLMM approach allowed us to explicitly control for center bias by including the center proximity for both viewing conditions (upright and inverted, Figure 1d, h) of each fixated and non-fixated region as both a fixed effect and as an interaction term.

We then computed two separate fixation location GLMMs. First, we fit a model to just the upright scene data. Specifically, whether a scene region was fixated (1) or not fixated (0) was modeled as a function of meaning, GBVS, and center proximity with subject and scene treated as full random effects. This served as a reference model that demonstrated how meaning, GBVS, and center proximity

are typically associated with attention when scenes are viewed in their common upright orientation. Then, we fit a second fixation location GLMM model to the full scene data (i.e., upright and inverted scene data) to estimate how inversion affected the association between attention and meaning, GBVS, and center proximity. Specifically, whether a region was fixated (1) or not fixated (0) was modeled as a function of the meaning, GBVS, center proximity, and viewing condition (dummy coded). Subject and scene were again treated as full random effects. The inversion interaction terms were of primary interest as they are the terms that reflect the effect of scene inversion on where viewers looked as it relates to meaning (semantic-guidance), GBVS (image-guidance), and center bias relative to the upright viewing condition.

	Fixed effects					Random effects, *SD*	
Predictors	β	95% CI	*SE*	z statistic	p	Subject	Scene
Intercept	-0.288	[-0.448 -0.127]	0.083	-3.46	0.001	0.123	0.776
Meaning	1.900	[1.721 2.078]	0.091	13.00	$< .001$	0.138	0.831
GBVS	0.510	[0.321 0.698]	0.096	5.33	$< .001$	0.225	0.828
Center Proximity	0.666	[0.442 0.889]	0.114	5.86	$< .001$	0.384	0.874
Meaning:GBVS	-0.108	[-0.271 0.054]	0.083	-1.31	0.191	0.127	0.706
Meaning:Center Proximity	0.159	[-0.068 0.386]	0.116	1.37	0.171	0.061	1.076
GBVS:Center Proximity	0.019	[-0.155 0.193]	0.089	0.22	0.829	0.225	0.775
Meaning:GBVS:Center Proximity	0.167	[-0.015 0.349]	0.093	1.78	0.075	0.103	0.857

Table 1: Fixation location general linear mixed effects model results: upright only model. *Beta estimates (β), 95% confidence intervals (CI), standard errors (SE), $Z-$statistic, and p-values (p) for each fixed effect and standard deviations* (SD) *for the random effects of subject and scene.*

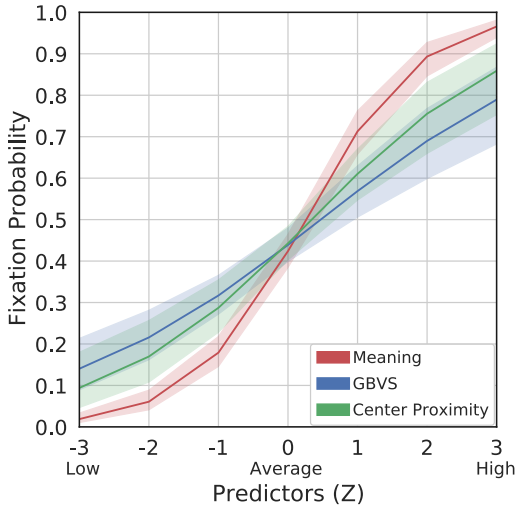

Figure 2. Effects of meaning, GBVS, and center proximity: upright only model. The line plots show the fixed effects of meaning, image saliency, and center proximity as a function of fixation probability. A scene region's meaning value showed the strongest association with attention, followed by center proximity, and image salience. Error bands reflect 95% confidence intervals.

2.4.3. Fixation duration general linear mixed effects models. Finally, we used a GLMM to examine how fixation durations changed with inversion as a function of meaning, GBVS, and center proximity values. Specifically, we modeled fixation duration using a GLMM with a gamma distribution and identity link function (Lo & Andrews, 2015) in the *lme4* package (Bates et al., 2015) as a function of meaning, GBVS, center proximity, and viewing condition. Just like in the fixation location model, condition was dummy coded and subject and scene were treated as full random effects. The fixation duration model was used to offer some additional insights into how processing of semantic features, uninterpreted image features, and center bias are altered by scene inversion relative to upright viewing.

Results

Before examining scene inversion effects, it is helpful to first show how attention typically varies as a function of a scene region's meaning, GBVS, and center proximity value in upright scene viewing only. Table 1 and Fig. 2 show the general linear mixed effects model (GLMM) results for the upright viewed scene data only. The model results showed significant positive fixed effects of meaning ($\beta = 1.90$, CI $[1.72, 2.07]$, $p < .001$), GBVS ($\beta = 0.51$, CI $[0.32, 0.69]$, $p < .001$), and center proximity ($\beta = 0.66$, CI $[0.44, 0.88]$, $p < .001$). No significant interactions were observed. The fixed effects are shown as a function of fixation probability in Fig. 2, with meaning showing the greatest effect on the probability a scene region would be fixated followed by center proximity and GBVS image salience. This pattern of results replicates previous findings that show a stronger effect of local meaning on attentional guidance than local image salience (for review see Henderson, Hayes, Peacock, & Rehrig, 2019). Finally, the random effects revealed larger scene variability than subject variability consistent with previous findings (Nuthmann et al., 2017; Henderson & Hayes, 2017; Hayes & Henderson, 2021b). With these typical effects in mind, we can now examine how these effects are altered by scene inversion.

The full (upright and inverted) fixation location GLMM results are shown in Table 2 and Fig. 3 and reflect how attention to each feature was affected by scene inversion (i.e., the difference in slope between the upright baseline condition and scene inverted condition). The model results indicated a negative meaning by inversion interaction ($\beta = -1.18$, CI $[-1.38, -0.98]$, $p < .001$). Fig. 3b shows a plot of the model interaction effect as a function of fixation probability and meaning value by condition. In the upright scene viewing condition, the probability of a scene region being fixated increased strongly as the meaning value increased. However, in the inverted viewing condition a region's meaning value was essentially uninformative. That is, high meaning scene regions were no more likely to be fixated than low meaning scene regions. This finding suggests that scene inversion strongly disrupts local semantic guidance.

As a comparison, we also examined the effect of scene inversion on low-level, presemantic image saliency using the GBVS maps. We observed a smaller negative GBVS by inversion interaction ($\beta = -0.21$, CI $[-0.37, -0.05]$, $p = .009$). The inversion effect on image salience is shown visually as a function of fixation probability in Fig. 3c for the upright and inverted scene conditions, suggesting that image guidance remained largely intact during inverted scene viewing. These findings support the view that low-level image features are predominantly presemantic and attentionally distinct from semantic features like local semantic density.

The model also revealed a positive center proximity by inversion interaction ($\beta = 0.60$, CI $[0.46, 0.74]$, $p < .001$). The model center proximity inversion effect is shown visually in Fig. 3d by viewing condition, indicating that when scenes were inverted viewers were more likely to fixate more central scene regions and less likely to fixate more peripheral scene regions compared to the upright viewing condition (Fig. 3d). These findings suggest that observer center bias is at least partially modulated by scene semantics, since the center of the scene and visual information it contained were constant across the upright and inverted viewing conditions.

In addition to the significant inversion effects, we also observed a significant negative meaning by GBVS interaction and positive meaning by center proximity interaction. As shown in Fig. 3e and

	Fixed effects					Random effects, *SD*	
Predictors	β	95% CI	*SE*	z statistic	p	Subject	Scene
Intercept	-0.254	[-0.405 -0.103]	0.077	-3.299	0.001	0.142	0.708
Meaning	1.220	[1.059 1.381]	0.082	14.841	$< .001$	0.124	0.759
GBVS	0.633	[0.467 0.800]	0.085	7.459	$< .001$	0.209	0.730
Center Proximity	0.654	[0.443 0.866]	0.108	6.061	$< .001$	0.412	0.789
Inversion	0.106	[-0.045 0.257]	0.077	1.380	0.167	0.045	0.732
Meaning:Center Proximity	0.193	[0.072 0.314]	0.062	3.121	0.002	0.118	0.518
GBVS:Center Proximity	0.075	[-0.028 0.179]	0.053	1.422	0.155	0.163	0.423
Meaning:GBVS	-0.290	[-0.410 -0.170]	0.061	-4.728	$< .001$	0.083	0.542
Meaning:Inversion	-1.183	[-1.380 -0.985]	0.101	-11.739	$< .001$	0.167	0.923
GBVS:Inversion	-0.213	[-0.372 -0.054]	0.081	-2.623	0.009	0.110	0.728
Center Proximity:Inversion	0.609	[0.469 0.749]	0.072	8.510	$< .001$	0.132	0.614

Table 2: *Fixation location general linear mixed effects model: full model.* Beta estimates (β), 95% confidence intervals (CI), standard errors (SE), $Z-$statistic, and p-values (p) for each fixed effect and standard deviations (*SD*) for the random effects of subject and scene.

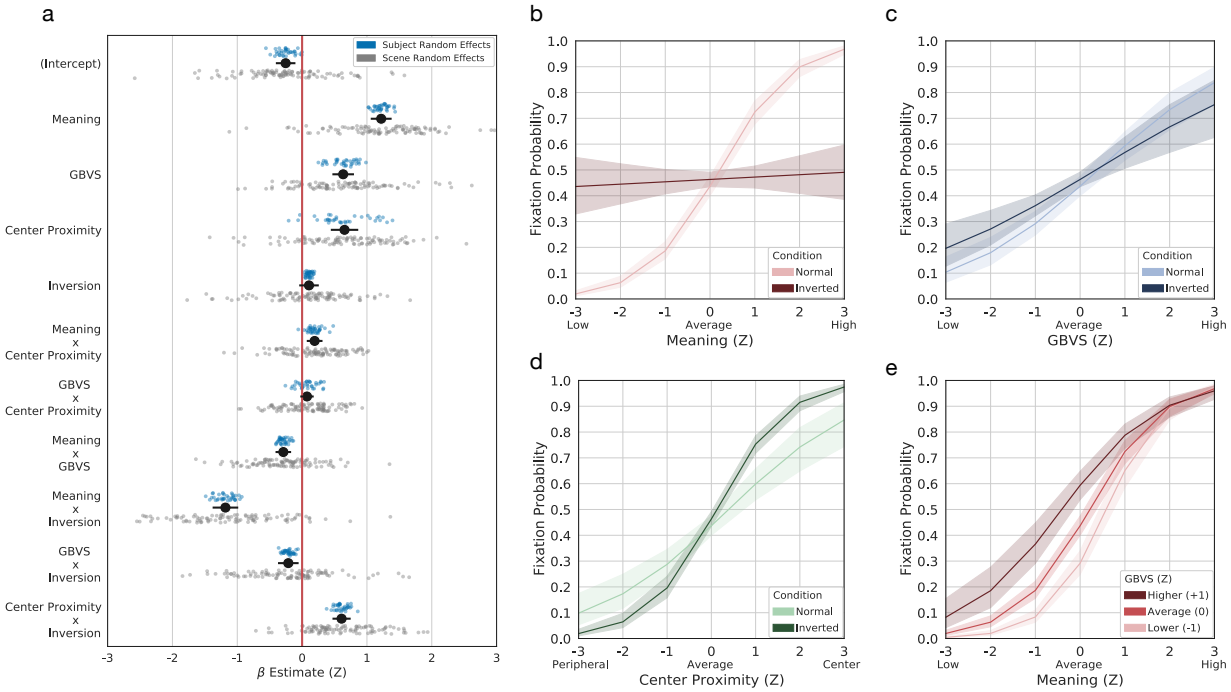

Figure 3. Fixation location general linear mixed effects model: full model. Whether a scene region was fixated or not was modeled as a function of its meaning map, GBVS image saliency map, center proximity map, and inversion condition. The black dots with lines show the beta weight estimates from the model and their 95% confidence intervals for each model term. Subject (blue dots) and scene (grey dots) were both accounted for in the model as full random effects. The line plots to the right show the interactions between inversion and meaning (b), image saliency (c), and center proximity (d). Panel e shows the marginal effects of meaning, GBVS, and center proximity. All error bands reflect 95% confidence intervals.

	Fixed effects					Random effects, *SD*	
Predictors	β	95% CI	*SE*	*t* statistic	*p*	Subject	Scene
Intercept	259.342	[252.738 265.947]	3.370	76.966	$< .001$	17.929	11.564
Meaning	4.658	[2.340 6.976]	1.183	3.938	$< .001$	4.001	2.135
GBVS	-1.546	[-5.206 2.114]	1.867	-0.828	0.408	5.889	9.075
Center Proximity	5.877	[1.624 10.129]	2.170	2.709	0.007	9.341	9.251
Inversion	4.110	[-0.706 8.926]	2.457	1.673	0.094	8.051	12.543
Meaning:Center Proximity	-1.591	[-5.106 1.924]	1.793	-0.887	0.375	5.466	8.556
GBVS:Center Proximity	4.178	[1.242 7.114]	1.498	2.789	0.005	4.909	8.195
Meaning:GBVS	-1.678	[-5.173 1.817]	1.783	-0.941	0.347	5.404	9.027
Meaning:Inversion	-9.701	[-14.931 -4.470]	2.669	-3.635	$< .001$	9.481	12.465
GBVS:Inversion	-0.103	[-6.008 5.802]	3.013	-0.034	0.973	7.509	17.972
Center Proximity:Inversion	2.937	[-2.370 8.244]	2.708	1.085	0.278	6.176	16.647

Table 3: Fixation duration general linear mixed effects model results. Beta estimates (β) in msec, 95% confidence intervals (CI), standard errors (SE), *t*statistic, and *p*-values (*p*) for each fixed effect and standard deviations (*SD*) for the random effects of subject and scene.

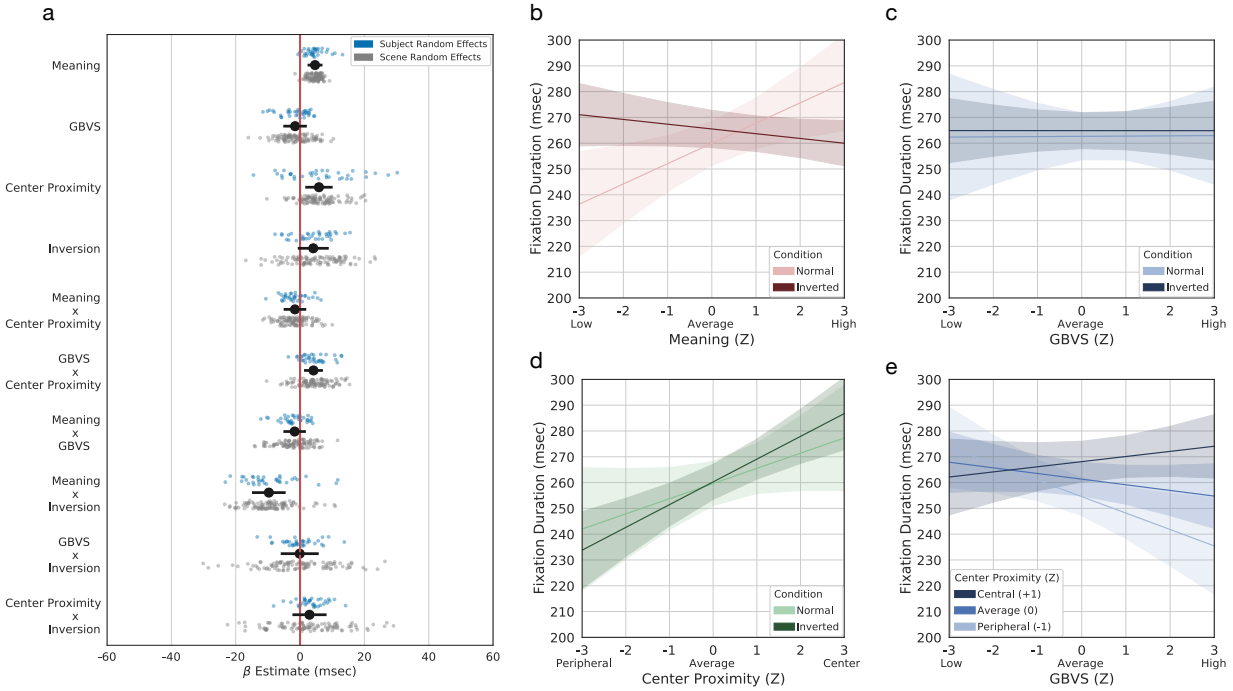

Figure 4. Fixation duration general linear mixed effects model. Fixation duration was modeled as a function of each fixated region's meaning map, GBVS image saliency map, center proximity map value, and inversion condition. The black dots with lines show the beta weight estimates from the model and their 95% confidence intervals for each model term. Subject (blue dots) and scene (grey dots) were both accounted for in the model as full random effects. The line plots to the right show the interactions between inversion and meaning (b), image saliency (c), and center proximity (d). Panel e shows the significant interaction between GBVS and center proximity. All error bands reflect 95% confidence intervals.

Table 2, the meaning by GBVS interaction ($\beta = -0.29$, CI $[-0.41, -0.17]$, $p < .001$) reflected that as a scene region's meaning value increased, image salience was increasingly discounted in the guidance of attention. This finding is consistent with previous work showing scene semantics can override image salience in the control of attention (Wu et al., 2014; Hwang et al., 2011; Malcolm et al., 2016; Hayes & Henderson, 2021b). The meaning by center proximity interaction ($\beta = 0.19$, CI $[0.07, 0.31]$, $p = .002$) was somewhat less interesting, reflecting that for very low and very high meaning regions, how close a region was to the center had a smaller affect on fixation probability.

Taken together, the upright only GLMM and full GLMM fixation location models show that local semantic density, image salience, and center bias are each uniquely impacted by scene inversion. In order to gain further insight into the underlying mechanisms that may contribute to these observed patterns, we also modeled how fixation durations were affected by scene inversion as a function of a scene region's meaning, GBVS, and center proximity value.

The fixation duration GLMM results are shown in Table 3 and Fig. 4. As can be seen, Table 3 has the same terms as Table 2, the only difference is here the dependent variable is fixation duration whereas before it was whether a region was fixated or not, reflected by the change in x-axis units in Fig. 4a y-axes in Fig. 4b, c, d, and e. Again, we observed different patterns of interaction between our predictors and scene inversion. In this case, only meaning showed a significant interaction with inversion ($\beta = -9.70$, CI $[-14.93, -4.47]$, $p < .001$) while GBVS and center proximity showed no significant inversion interaction (see Table 3 and Fig. 4c and d). The meaning interaction indicated that under upright scene viewing, fixation duration increased as a scene region's meaning value increased (Fig. 4b). However, the fixation duration meaning pattern reversed when scenes were inverted, with fixation durations decreasing as a region's meaning value increased. The dissociation with local meaning in upright and inverted viewing suggests that the attentional and cognitive processing mechanisms responsible for determining fixation duration are also significantly altered by scene inversion. Only one other significant interaction was observed in the fixation duration GLMM, a GBVS by center proximity interaction ($\beta = 4.17$, CI $[1.24, 7.11]$, $p = .005$). As shown in Fig. 4e, this interaction reflected a decrease in fixation duration for highly visually salient regions as distance from the center increased.

Together the fixation location GLMM and the fixation duration GLMM results suggest that scene inversion produces a unique pattern of changes in how semantic guidance, image guidance, and observer center bias are associated with where and for how long people look in real-world scenes.

3. Discussion

Semantic guidance and image guidance theories make different predictions for how attention will be allocated to different scene regions as exemplified by image salience maps (Harel et al., 2006) and semantic feature maps (Henderson & Hayes, 2017). Here we examined how the association between attention and semantically interpreted features, semantically uninterpreted low-level image features, and observer center bias changed between upright and inverted scene viewing. Our results showed distinct patterns of change when scenes were inverted. Semantic guidance, as indexed by local semantic density, was knocked completely offline when scenes were viewed from an inverted viewpoint, despite being the best predictor of where people looked in upright scenes. Surprisingly, when semantic guidance went offline, image salience did not fill the attentional void; instead, center bias did. Finally, we showed that fixation durations to local meaning were uniquely impacted by scene inversion. In upright viewing we observed larger fixation durations for more meaningful regions, but this pattern reversed for inverted scene viewing.

The effect of scene inversion on semantic guidance to local semantic density was striking. A scene region's meaning value went from being the strongest predictor of attention under upright viewing to being completely uninformative when scenes were inverted. Importantly, this finding was not specific

to a single isolated scene region or object, but was observed taking into account attention across the entire scene. This finding suggests that the mapping between local meaning and the current scene stimulus is strongly dependent on the scene stimulus matching our internal models of scene structures gained from experience, and without that, the attention system is not able to effectively leverage our stored semantic knowledge to guide attention.

In the absence of semantic guidance, image-guidance theory would predict observers should lean more heavily on image salience to guide attention. However, our data does not support this hypothesis. Image-guidance based on low-level presemantic image features remained relatively unaffected by scene inversion, actually showing a mild deficit instead of any enhancement when semantic guidance was disrupted. While the inversion effect was mild (See Fig. 3c) it was observed for all subjects and about 60% of scenes (See Fig. 3a random subject and scene slope estimates). The deficit in image saliency under scene inversion suggests that image saliency may be mildly modulated by changes in semantic guidance. This is another theoretically important finding, as it offers a new intriguing piece of evidence that scene semantics may have some influence on early visual processing mechanisms (Teufel, Dakin, & Fletcher, 2018).

The significant enhancement of center bias we observed with scene inversion was also unexpected. Given that the visual scene information remained at a consistent distance from the screen center in the upright and inverted viewing conditions, the naive prediction would be that center bias would be the same in the upright and inverted conditions. Instead, our findings suggest that the strength of center bias is not purely a function of the screen center location, the distance of image features from the screen center, or occulomotor regularities (Tatler, 2007), but is also partially modulated by how readily semantic scene content can be used to guide attention. That is, when semantic guidance was disrupted by scene inversion, observer center bias seemed to fill the attentional vacuum that was left rather than image guidance.

Why does attention to local semantic density change so drastically in the inverted scene condition, and why is observer center bias enhanced instead of image salience? Unlike previous scene inversion work, the disruption of attention to local semantic density cannot be directly attributed to a disruption in object-scene semantics because the meaning map ratings are for random, isolated scene patches without scene context. However, it is likely that when scene context is available during viewing, our knowledge of the scene category does help point us to scene regions that are more likely to be semantically rich (e.g., the counter top in a kitchen or a desk in an office). Therefore, it may be that the disruption to scene-object semantics *indirectly* affects the ability to guide attention to local semantic density. The fixation duration results also provide some potential clues. The decrease in fixation duration to semantically rich regions in the inversion condition suggests participants are actually being slightly repelled from processing semantically rich regions. This in conjunction with the increase in observer center bias we observed in the fixation location GLMM suggests that participants may be under substantial cognitive load in the inverted condition and seeking refuge by looking at less semantically rich regions. Therefore, one plausible explanation is that while participants are able to accurately estimate local meaning for isolated regions regardless of orientation, it is simply too cognitively taxing (and perhaps too slow as a result) to use local semantic density to effectively guide attention during scene viewing. This interpretation is consistent with previous findings showing that increased working memory load during scene viewing increases observer center bias (Cronin, Peacock, & Henderson, 2021).

While this study has shown a number of important findings, it also has limitations that should be addressed in future work. First, the current work used an active scene memorization viewing task and it may be that other active tasks (e.g., visual search) or no-task (e.g., free viewing) show different inversion effect patterns than we observed here. However, given the important role local semantic density has been shown to play in a wide variety of different viewing tasks (Henderson et al., 2019), we speculate that similar deficits may appear with scene inversion in other tasks. Second, the use of

local meaning maps (Henderson & Hayes, 2017, 2018), is only one type of semantic feature map. It would also be useful in future work to examine other types of semantic feature maps (e.g., graspability and reachability Rehrig et al., 2020, or object-object and scene-object semantic similarity Hayes & Henderson, 2021b) to determine if the inversion effects we observed here generalize to other types of semantic features. Finally, it would be useful to replicate the negative image saliency inversion effect in other low-level image saliency models to verify that this effect is not specific to the GBVS model.

In summary, we quantified how scene inversion impacts attention to semantically interpreted features, uninterpreted image features, and observer center bias for the first time. We found that scene inversion affected each in a unique way: local semantic guidance was knocked offline, image guidance was mildly impaired, and observer center bias was enhanced. In addition, an analysis of the effect of scene inversion on fixation durations suggested that observers may be actively repelled from semantically rich regions when viewing inverted scenes. These findings reinforce that image guidance and semantic guidance are attentionaly distinct, provide novel evidence that observer center bias can be modulated by changes in semantic guidance, and offer tantalizing new clues for why semantic guidance is disrupted by inversion. More broadly, our results provide important new constraints for theories and computational models of attention by providing unique patterns of disruption and enhancement that should be observed for inverted scenes.

Supplementary material

`https://osf.io/bnwxv/`

Acknowledgements

This research was partially funded by BCS2019445 from the US National Science Foundation. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors declare no competing financial interests.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. doi: 10.18637/jss.v067.i01

Biederman, I. (1972). Perceiving real-world scenes. *Science*, *177*, 77-80.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*, 143-177.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185-207.

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.

Boyce, S. J., Pollatsek, A., & Rayner, K. (1989). Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception & Performance*, *15*, 556–566.

Buswell, G. T. (1935). *How people look at pictures*. Chicago: University of Chicago Press.

Castelhano, M. S., & Heaven, C. (2011). Scene context influences without scene gist: Eye movements guided by spatial associations in visual search. *Psychonomic Bulletin & Review*, *18*, 890-896.

Castelhano, M. S., Mack, M., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, *9*, 1-15.

Cronin, D. A., Peacock, C. E., & Henderson, J. M. (2021). Visual and verbal working memory loads interfere with scene-viewing. *Attention, Perception, & Psychophysics*, *82*, 2814-2820.

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 of the United States of America*, *116*, 11687-11692.

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*(2), 1-19.

Epstein, R. A., Higgins, J. S., Parker, W., Aquirre, G. K., & Cooperman, S. (2006). Cortical correlates of face and scene inversion: A comparison. *Neuropsychologia*, *47*(7), 1145–1158.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In *Proceedings of the 19th international conference on neural information processing systems* (p. 545-552). Cambridge, MA, USA: MIT Press.

Hayes, T. R., & Henderson, J. M. (2019). Scene semantics involuntarily guide attention during visual search. *Psychonomic Bulletin and Review*, *26*(5), 1683-1689.

Hayes, T. R., & Henderson, J. M. (2020). Center bias outperforms image salience but not semantics in accounting for attention during scene viewing. *Attention, Perception, & Psychophysics*, *82*(3), 985-994.

Hayes, T. R., & Henderson, J. M. (2021a). Deep saliency models learn low-, mid-, and high-level features to predict scene attention. *Scientific Reports*, *11*, 1-13.

Hayes, T. R., & Henderson, J. M. (2021b). Looking for semantic similarity: What a vector space model of semantics can tell us about attention in real-world scenes. *Psychological Science*, *32*, 1262–1270.

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. (2011). Eye movements and scene perception. In I. S. P. Liversedge, D. Gilchrist, & S. Everling (Eds.), *The oxford handbook of eye movements* (p. 593-606). Oxford University Press.

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 R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movements: A window on mind and brain* (p. 537-562). Oxford University Press.

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., Peacock, C. E., & Rehrig, G. (2019). Meaning and attentional guidance in scenes: A review of the meaning map approach. *Vision*, *2*(19), 1-10.

Henderson, J. M., Hayes, T. R., Peacock, C. E., & Rehrig, G. (2021). Meaning maps capture the density of local semantic features in scenes: A reply to Pedziwiatr, Kummerer, Wallis, Bethge & Teufel (2021). *Cognition*, *214*, 104742.

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.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*, 850-856.

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

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of experimental psychology. General*, *127*, 398-415.

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.

Hwang, A. D., Wang, H. C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research*, *51*(10), 1192-1205.

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.

Jolicoeur, P. (1988). Mental rotation and the identification of disoriented objects. *Canadian Journal of Psychology*, *42*(4), 461-478.

Kelley, T. A., Chun, M. M., & Chua, K. P. (2003). Effects of scene inversion on change detection target matched for visual salience. *Journal of Vision*, *3*(1), 1-5.

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

Lo, S., & Andrews, S. (2015). To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1171.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology*, *4*, 565-572.

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

Malcolm, G. L., Groen, I. I. A., & Baker, C. I. (2016). Making sense of real-world scenes. *Trends in Cognitive Science*, *20*, 843-856.

Nuthmann, A., Einhäuser, W., & Schütz, I. (2017). How well can saliency models predict fixation selection in scenes beyond center bias? A new approach to model evaluation using generalized linear mixed models. *Frontiers in Human Neuroscience*, *11*, 491.

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. (2019a). Meaning guides attention during scene viewing even when it is irrelevant. *Attention, Perception, & Psychophysics*, *81*, 20-34.

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019b). The role of meaning in attentional guidance during free viewing of real-world scenes. *Acta Psychologica*, *198*, 1-8.

Peterson, M., & Gibson, B. (1994). Must Figure-Ground Organization Precede Object Recognition? An Assumption in Peril. *Psychological Science*, *5*, 253 - 259.

Potter, M. (1975). Meaning in visual search. *Science*, *187*, 965-966.

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rehrig, G., Peacock, C. E., Hayes, T. R., Henderson, J., & Ferreira, F. (2020). Where the action could be: Speakers look at graspable objects and meaningful scene regions when describing potential actions. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *46*(9), 1659-1681.

Rock, I. (1974). The perception of disoriented figures. *Scientific American*, *230*, 78-85.

Rock, I., & DiVita, J. (1987). A case of viewer-centered object perception. *Cognitive Psychology*, *19*(2), 280-293.

Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of Vision*, *7*(16), 1-20.

Shore, D. I., & Klein, R. M. (2000). The effects of scene inversion on change blindness. *Journal of General Psychology*, *127*, 27-43.

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.

Tarr, M. J., & Pinker, S. (1990). When does human object recognition use a viewer-centered reference frame? *Psychological Science*, *1*(4), 253-256.

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.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5), 1-23.

Teufel, C., Dakin, S. C., & Fletcher, P. C. (2018). Prior object-knowledge sharpens properties of early visual feature-detectors. *Scientific Reports*, *8*, 10853.

Võ, M. L. H., Boettcher, S. E. P., & Draschkow, D. (2019). Reading scenes: how scene grammar guides attention and aids perception in real-world environments. *Current opinion in psychology*, *29*, 205-210.

Võ, M. L. H., & Henderson, J. M. (2011). Object-scene inconsistencies do not capture gaze: evidence from the flash-preview moving-window paradigm. *Attention, Perception & Psychophysics*, *73*, 1742-1753.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*, 1395-1407.

Williams, C. C., & Castelhano, M. S. (2019). The Changing Landscape: High-level Influence on Eye Movement Guidance in Scenes. *Vision*, *3*(3), 33.

Wu, C. C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology*, *5*, 1-13.

Yarbus, A. L. (1967). *Eye movements and vision.* New York: Plenum.