Attention, Perception, & Psychophysics (2019) 81:20–34
https://doi.org/10.3758/s13414-018-1607-7

# Meaning guides attention during scene viewing, even when it is irrelevant

Candace E. Peacock[1,2] • Taylor R. Hayes[1] • John M. Henderson[1,2]

Published online: 23 October 2018

## Abstract

During real-world scene viewing, humans must prioritize scene regions for attention. What are the roles of low-level image salience and high-level semantic meaning in attentional prioritization? A previous study suggested that when salience and meaning are directly contrasted in scene memorization and preference tasks, attentional priority is assigned by meaning (Henderson & Hayes in Nature Human Behavior, 1, 743–747, 2017). Here we examined the role of meaning in attentional guidance using two tasks in which meaning was irrelevant and salience was relevant: a brightness rating task and a brightness search task. Meaning was represented by meaning maps that captured the spatial distribution of semantic features. Meaning was contrasted with image salience, represented by saliency maps. Critically, both maps were represented similarly, allowing us to directly compare how meaning and salience influenced the spatial distribution of attention, as measured by fixation density maps. Our findings suggest that even in tasks for which meaning is irrelevant and salience is relevant, meaningful scene regions are prioritized for attention over salient scene regions. These results support theories in which scene semantics play a dominant role in attentional guidance in scenes.

**Keywords** Attention · Scene perception · Eye movements · Meaning · Salience

Because we can only attend to a small portion of the visual information available to us, we have to select some regions of the visual scene for preferential analysis, at the expense of others, via attention. It is therefore important to understand the mechanisms by which we guide our attention through real-world scenes. A good deal of work on attentional guidance in scenes has focused on the idea that attention is driven by bottom-up, low-level image features, such as color, luminance, and edge orientation, that are combined into saliency maps (Borji, Parks, & Itti, 2014; Borji, Sihite, & Itti, 2013; Harel, Koch, & Perona, 2006; Itti & Koch, 2001). Saliency maps are appealing because they are computationally tractable and neurobiologically plausible (Henderson, 2007, 2017).

At the same time, there is strong evidence that visual attention is influenced by cognitive factors such as the semantic informativeness of objects and entities within a scene (Antes, 1974; Henderson, 2017; Henderson, Brockmole, Castelhano, & Mack, 2007; Mackworth & Morandi, 1967), along with the viewer's task and current goal (Buswell, 1935; Hayhoe & Ballard, 2005; Hayhoe, Shrivastava, Mruczek, & Pelz, 2003; Henderson, 2007, 2017; Henderson & Hollingworth, 1999; Navalpakkam & Itti, 2005, 2007; Rothkopf, Ballard, & Hayhoe, 2016; Tatler, Hayhoe, Land, & Ballard, 2011; Yarbus, 1967). Yet much of the research on attentional guidance has continued to focus solely on image salience. One reason for the popularity of image salience is that it is relatively straightforward to compute and represent. In contrast, it has been less clear how to generate and represent the spatial distribution of semantic features across a scene. To directly compare image salience to semantic informativeness, it is necessary to represent scene meaning in a format equivalent to that of image salience.

To address this issue, we have recently introduced the concept of meaning maps as a way to represent the spatial distribution of scene semantics (Henderson & Hayes, 2017). To generate meaning maps, Henderson and Hayes (2017) used crowd-sourced responses in which naïve participants rated the meaning of image patches from real-world scenes. Specifically, photographs of scenes were divided into a dense

✉ Candace E. Peacock
cepeacock@ucdavis.edu

[1] Center for Mind and Brain, University of California, Davis, CA, USA

[2] Department of Psychology, University of California, Davis, CA, USA

array of objectively defined circular overlapping patches at coarse and fine spatial scales. These patches were then shown to raters, who rated how informative or recognizable each patch was (see also Antes, 1974; Mackworth & Morandi, 1967). Finally, meaning maps of each scene were created by interpolating the ratings at each spatial scale and averaging across the two scales.

Meaning maps provide a pixel-by-pixel prediction of semantic content across a scene, just as saliency maps provide a pixel-by-pixel prediction of salience across a scene. Since meaning maps are represented in the same format as saliency maps, their predictions for visual attention can be directly compared to those from saliency maps, by using the methods that have typically been used to compare the relationship between saliency maps and attention (Carmi & Itti, 2006; Itti, Koch, & Niebur, 1998; Parkhurst, Law, & Niebur, 2002; Torralba, Oliva, Castelhano, & Henderson, 2006). In this way, meaning maps and saliency maps together provide a way to compare how meaning and salience influence visual attention during real-world scene viewing.

Henderson and Hayes (2017) investigated the degrees to which meaning maps and saliency maps predicted visual attention in real-world scenes during memorization and aesthetic judgment tasks. In that study, attention maps were created on the basis of the locations of eye fixations. The results showed that meaning maps and saliency maps were highly correlated, and that both were able to predict the spatial distribution of attention in scenes. Importantly, in both tasks meaning accounted for significantly more of the variance in attention than did image salience. Furthermore, when the variance due to salience was controlled, meaning accounted for significantly more of the remaining variance in attention, but when meaning was controlled, no additional variance in attention was accounted for by salience. These results held across the entire viewing time. Henderson and Hayes (2018) replicated this pattern of results using attention maps constructed from duration-weighted fixations, and Henderson, Hayes, Rehrig, and Ferreira (2018) showed that the results extended to scene description tasks. In total, the findings showed that meaning (rather than image salience) was the main driver of visual attention.

Although the data favoring meaning over image salience have been clear, it could be argued that the viewing tasks used to compare meaning and image salience were biased toward meaning. That is, it might be that memorization, aesthetic preference, and scene description tasks require the viewer to focus on the semantic features of scenes. If this is true, then it may be that the advantage for meaning over salience is restricted to viewing tasks that specifically require analysis of meaning. To address this hypothesis, in the present study we investigated whether attention continues to be guided by meaning during scene viewing, even when salience is relevant and meaning is irrelevant to the viewer's task.

Specifically, in the present study we used two tasks that were designed to emphasize salience and to eliminate the need for meaning in attentional guidance: a brightness rating task, in which participants rated scenes for overall brightness, and a brightness search task, in which participants counted the number of bright patches within scenes (Fig. 1). Critically, these tasks were designed to make meaning task-irrelevant and salience task-relevant. If the use of meaning to guide attention is task-based, then the relationship between meaning and attention found in our earlier studies should no longer be observed in these conditions. On the other hand, if the use of meaning to guide attention during scene viewing is a fundamental property of the attention system, then we should continue to observe a relationship between meaning and attention even when only salience is relevant to the task.

# Method

## Eye-tracking

**Participants** Thirty University of California, Davis, undergraduate students with normal or corrected-to-normal vision participated in the experiment (25 females, five males; average age = 20.84 years). All participants were naïve concerning the purpose of the experiment and provided verbal consent. The eye movement data from each participant were filtered for excessive track losses due to blinks or loss of calibration. Following Henderson and Hayes (2017), we averaged the percent signal ([number of good samples/total number of samples] × 100) for each trial and participant using custom MATLAB code. The percent signal for each trial was then averaged for each participant and compared to an a priori 75% criterion for signal. Overall, all participants had greater than 75% signal, resulting in no removed participants.

**Apparatus** Eye movements were recorded using an EyeLink 1000+ tower-mounted eyetracker (spatial resolution 0.01° rms) sampling at 1000 Hz (SR Research, 2010). Participants sat 85 cm away from a 21-in. monitor, so that the scenes subtended approximately 26.5° × 20° of visual angle at 1,024 × 768 pixels. Head movements were minimized by using a chin and forehead rest. Although viewing was binocular, eye movements were recorded from the right eye. The experiment was controlled with the SR Research Experiment Builder software (SR Research, 2010).

**Stimuli** The stimuli consisted of 40 digitized photographs (1, 024 × 768 pixels) of indoor and outdoor real-world scenes. The scenes were luminance-matched across the scene set by converting the RGB image of the scene to LAB space and scaling the luminance channels of all scenes from 0 to 1. All

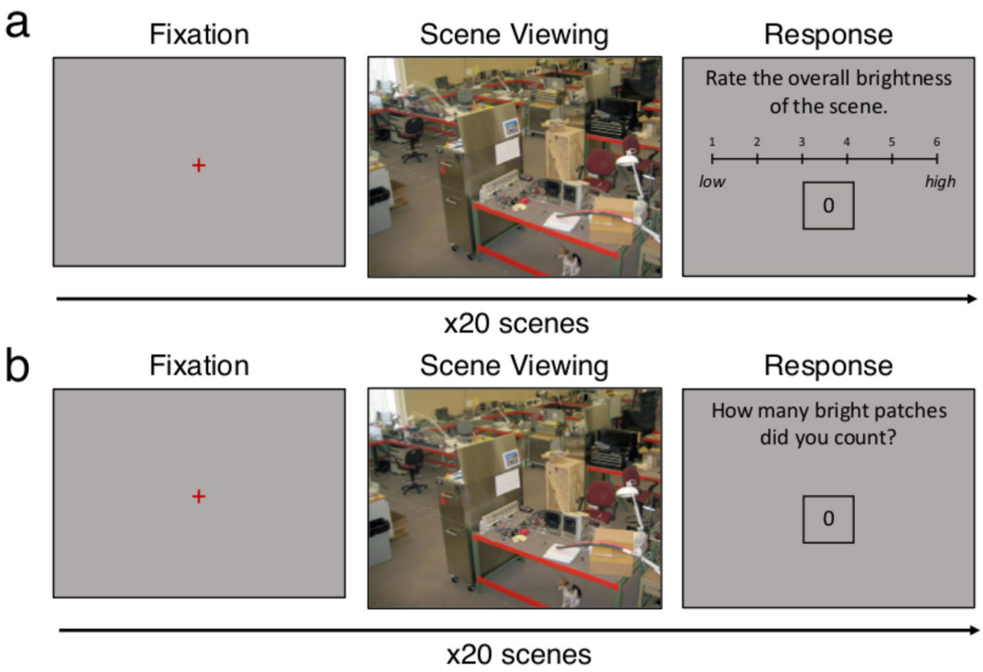


**Fig. 1** Trial structure for the two tasks: The trial structures for (a) the brightness rating task and (b) the brightness search task

instruction, calibration, and response screens were luminance-matched to the average luminance ($M = 0.45$) of the scenes.

**Procedure** Each participant completed two scene-viewing conditions in a within-subjects design: a brightness rating task and a brightness search task (Fig. 1). During the brightness rating task, participants were instructed to rate the overall brightness of the scene on a scale from 1 to 6 (1 = *very dark*, 2 = *dark*, 3 = *somewhat dark*, 4 = *somewhat bright*, 5 = *bright*, and 6 = *very bright*). During the brightness search task, participants were instructed to count the number of bright patches within the scene. Because the goal of this study was to assess whether we could eliminate the relationship between meaning and attention in tasks that did not require the use of meaning, we emphasized speed and accuracy during both tasks. Participants were given a maximum scene-viewing time of 12 s (as had been done in Henderson & Hayes, 2017), but they had the option to terminate the scene and continue to the response screen earlier by pressing a key on a button box (RESPONSEPixx; VPixx Technologies, Saint-Bruno, CA). We included the early-termination option so that they could focus on task-relevant eye movement behavior. Following their button press or the maximum of 12 s of scene presentation, participants were shown a response screen in which the number 0 was enclosed in a square (Fig. 1). Then, participants used left and right buttons on the button box to, respectively, increase or decrease the value of the number until it matched their rating or patch count for that scene. They then pressed the center key to continue to the next scene.

Before starting the experiment, participants completed two practice trials in which they were familiarized with each condition and the button-box. After the practice trials, a 13-point calibration procedure was performed 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°. The presentation of each scene was preceded by a drift correction procedure, and the eye-tracker was recalibrated when the calibration was not accurate. The calibration was also repeated between task blocks.

The 40 scene stimuli were randomly divided into two scene sets (Set A and Set B), each composed of 20 scenes, and for each participant, each set was assigned to one task. Task order and scene set assignment was fully counterbalanced across all participants. Additionally, the scenes within each set were presented in a randomized order for each participant in each condition.

## Analysis

All analyses were chosen a priori and on the basis of our previous work (Henderson & Hayes, 2017, 2018; Henderson et al., 2018).

### Data segmentation and outliers

Fixations and saccades were segmented with EyeLink's standard algorithm using velocity and acceleration thresholds (30°/s and 9500°/s$^2$; SR Research, 2010). The eye movement data were imported offline into Matlab using the EDFConverter tool. The first fixation in each scene, always located at the center of the display as a result of the pretrial fixation marker, was eliminated from the analysis.

Additionally, any fixations that were shorter than 50 ms and longer than 1,500 ms were eliminated, as outliers. This outlier removal process resulted in the loss of 3.94% of the data.

## Attention maps

Attention maps were generated as described in Henderson and Hayes (2017). Briefly, a fixation frequency matrix based on the locations (*x*, *y* coordinates) of all fixations was generated across participants for each scene. A Gaussian low-pass filter with a circular boundary and a cutoff frequency of - 6 dB was applied to each matrix, to account for foveal acuity and eye-tracker error (Fig. 2). The spatial extent of the low-pass filter was 236 pixels in diameter.

## Meaning maps

Meaning maps were generated as per Henderson and Hayes (2017). Because the nature of our tasks resulted in peripheral fixations, we used both unbiased and center-biased meaning maps (Fig. 2). Overall, the unbiased maps provided better predictive power than the center-biased maps. However, we included analyses from both, because center-biased maps are standard in the literature and thus provide a basis for comparison with previous studies. The center-biased meaning maps were generated by applying a multiplicative center bias operation to the meaning maps using the same center bias present in the saliency maps.

**Participants** Scene patches were rated by 165 participants on Amazon Mechanical Turk. These participants were recruited from the United States, had a HIT (human intelligence task) approval rate of 99% and 500 HITs approved, and were only allowed to participate in the study once. The participants were paid $0.50 cents per assignment, and all participants provided informed consent.

**Stimuli** The stimuli consisted of the 40 digitized photographs used in the present experiment. Each scene was decomposed into a series of partially overlapping and tiled circular patches at coarse and fine spatial scales. The full patch stimulus set consisted of 12,000 unique fine patches and 4,320 unique coarse patches, for a total of 16,320 scene patches.

**Procedure** Each participant rated 300 random scene patches extracted from the scenes. Participants were instructed to assess the meaningfulness of each patch on the basis of how informative or recognizable they thought it was. During the instruction period, participants were provided with examples of two low-meaning and two high-meaning scene patches, to make sure they understood the task. They then rated the meaningfulness of the test patches on a 6-point Likert scale (*very low*, *low*, *somewhat low*, *somewhat high*, *high*, and *very high*). The patches were presented in random order and without scene context, so the ratings were based on context-independent judgments. Each unique patch was rated three times by three independent raters, for a total of 48,960 ratings. However, owing to the high degree of overlap across patches, each fine patch contained rating information from 27 independent raters, and each coarse patch from 63 independent raters.

Meaning maps were generated from the ratings by averaging, smoothing, and combining the fine and coarse maps from the corresponding patch ratings. The ratings for each pixel at each scale in each scene were averaged, producing average fine and coarse rating maps for each scene. The average fine and coarse rating maps were then smoothed using thin-plate

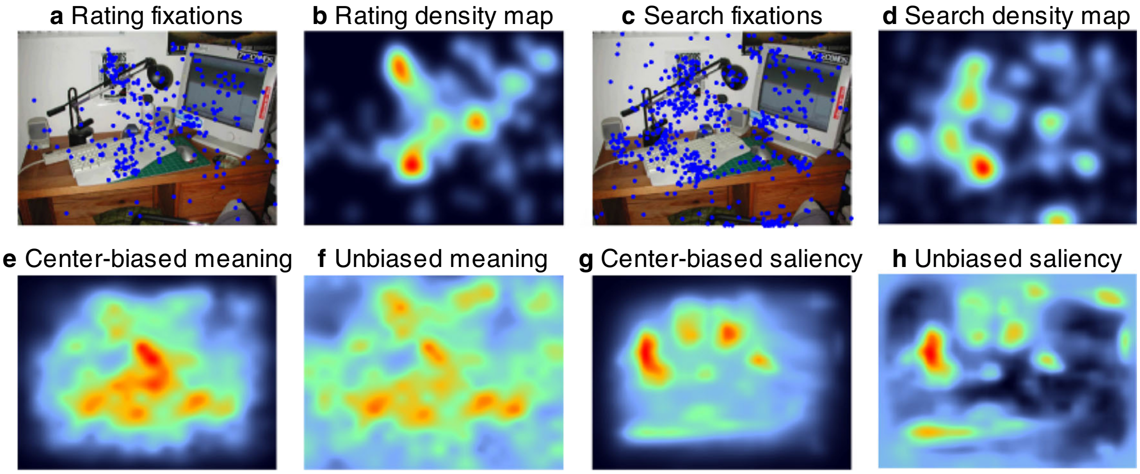


**Fig. 2** An example scene with the associated maps for each task. Panel a is an example scene, with fixation locations from all participants in the rating task aggregated and overlaid. Panel b is the fixation density map representing the example scene and fixation locations for the rating task. Panel c is the example scene with the fixations from the search task overlaid, and panel d is the fixation density map representing the example scene and fixation locations for the search task. Panel e is the center-biased meaning map, and panel f is the unbiased meaning map for the example scene. Panel g is the center-biased saliency map, and panel h is the unbiased saliency map for the example scene

spline interpolation based at the center of each patch (MATLAB "fit" function using the "thinplateinterp" method). Finally, the smoothed fine and coarse maps were averaged to produce the final meaning map for each scene.

### Saliency maps

The saliency maps for each scene were computed using the Graph-Based Visual Saliency (GBVS) toolbox with default settings (Harel et al., 2006). GBVS is a prominent salience model that combines maps of low-level image features in order to create image-based saliency maps (Fig. 2).

Center bias is a natural feature of GBVS saliency maps. To compare them to the unbiased meaning maps, we also generated GBVS maps without center bias (Fig. 2). These maps were created using a whitening method (Rahman & Bruce, 2015), a two-step normalization approach in which each saliency map is normalized to have 0 mean and unit variance. After this, a second, pixel-wise normalization is performed so that each pixel location across all the saliency maps has 0 mean and unit variance.

**Histogram matching** Following Henderson and Hayes (2017), the meaning and saliency maps were normalized to a common scale using image histogram matching, with the fixation density map for each scene serving as the reference image for the corresponding meaning and saliency maps. This was accomplished by using the Matlab function "imhistmatch" from the Image Processing Toolbox.

## Results

### Task comparisons

**Scene viewing** Because we gave participants the option to terminate each presentation trial early, we began by comparing the average scene-viewing (from scene onset to response) times for each scene during each condition, as well as the numbers of fixations per scene in each task (Fig. 3). The average scene-viewing time for the brightness rating task was 5,262.55 ms (*SD* = 3,141.39), with 15.56 fixations (*SD* = 9.84), and the average time for the brightness search task was 10,726.52 ms (*SD* = 2,420.55), with 32.28 fixations (*SD* = 8.20). Because the distributions were not normal (Fig. 3), Wilcoxon rank sum tests were conducted and showed that the scene-viewing times and numbers of fixations were significantly different between the rating and search tasks: *Z*s > 5.50, *p*s < .001. These results showed that participants tended to view scenes during the rating task for shorter durations than during the search task, with participants being much more likely to use the entire 12 s in the search than in the rating task. The finding that the rating task produced significantly shorter viewing durations than the search task suggests that participants only viewed the scenes for the amount of time necessary to complete each task. Given that the viewing times and numbers of fixations were very different between the tasks, we treated the two tasks separately in the following analyses.

**Response agreement** To verify that participants were staying on task and attending to brightness during the study, we examined response agreement in the rating and search tasks. If participants were on-task, then their responses should vary as a function of scene and be consistent within scenes. That is, participants should generally agree in their judgments of brightness in the rating task and in the number of bright regions in the search task. On the other hand, if participants were simply attending to scene content rather than following the instructions, then responses should be unsystematic across scenes and participants. As can be seen in Fig. 4, the former was true, suggesting that participants were indeed following the instructions.

### Overall scene analyses

Following Henderson and Hayes (2017), we used squared linear and semipartial correlations to quantify the degrees to which the meaning maps and saliency maps accounted for shared and unique variance in the attention maps. Specifically, we conducted two-tailed, two-sample *t* tests for the correlations across scenes in order to statistically compare the relative abilities of meaning and salience to predict attentional guidance.

For comparison to the literature, we tested how well traditional center-biased meaning and saliency maps could account for attention. In addition, because center bias was substantially reduced in the brightness search task as compared to the brightness rating task (Fig. 5), we also conducted analyses using unbiased meaning and saliency maps that excluded center bias.

**Brightness rating task** Using the center-biased maps, for squared linear correlations on average across all 40 scenes, meaning accounted for 55% of the variance in fixation density (*M* = .55, *SD* = .12), and salience accounted for 33% of the variance in fixation density (*M* = .33, *SD* = .14) (Fig. 6a). This difference between the meaning and saliency maps was significant: *t*(78) = 7.31, *p* < .001, 95% CI = [0.16, 0.28]. Similarly, for squared semipartial correlations, meaning accounted for 24% of the variance in fixation density (*M* = .24, *SD* = .13) after controlling for salience, but salience accounted for only 3% of the variance in fixation density after controlling for meaning (*M* = .03, *SD* = .03) (Fig. 6b). This difference was again significant: *t*(78) = 10.57, *p* < .001, 95% CI = [0.17, 0.25]. This pattern of results did not change when

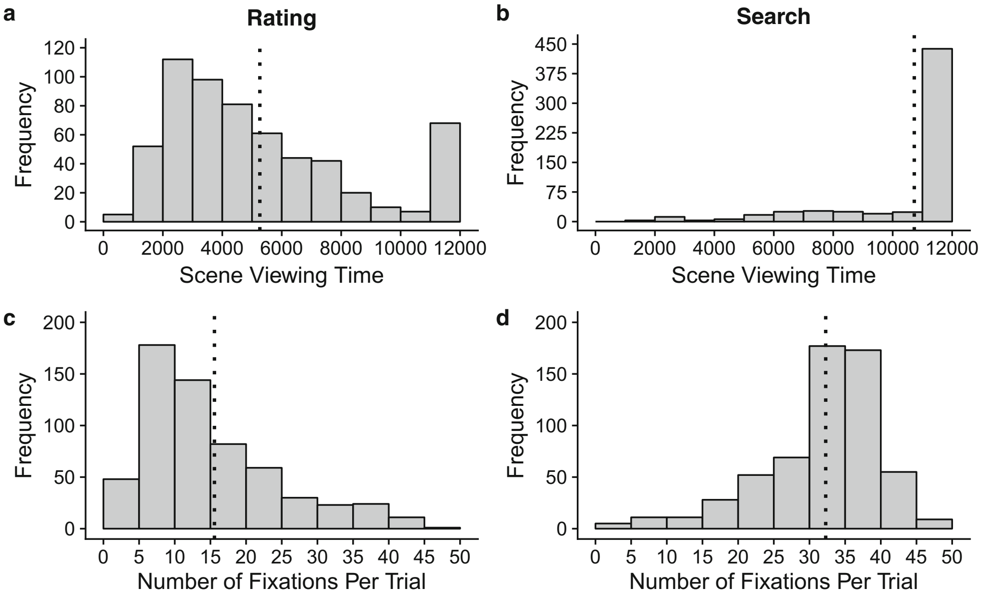


**Fig. 3** Scene-viewing times and numbers of fixations per trial for the brightness rating and brightness search tasks. Distributions are shown for the scene viewing times of (a) the brightness rating and (b) brightness search tasks, and the numbers of fixations per trial in (c) the brightness rating and (d) brightness search tasks. Each black, dotted vertical line represents the mean for each task

using the unbiased meaning and saliency maps: linear, $t(78) = 8.79$, $p < .001$, 95% CI = [0.16, 0.25]; semipartial, $t(78) = 9.62$, $p < .001$, 95% CI = [0.16, 0.25] (Figs. 6c and d). These findings suggest that meaning played a dominant role in the guidance of attention, even though meaning was irrelevant and salience was central to the brightness rating task.

**Brightness search task** Using the center-biased maps, meaning accounted for 22% of the variance in fixation density ($M$ = .22, $SD$ = .13), and salience accounted for 24% of the variance in fixation density ($M$ = .24, $SD$ = .12) (Fig. 7a). This difference was not significant: $t(78) = - 0.33$, $p = .74$, 95% CI = [-0.07, 0.05]. Similarly, for the semipartial correlations,

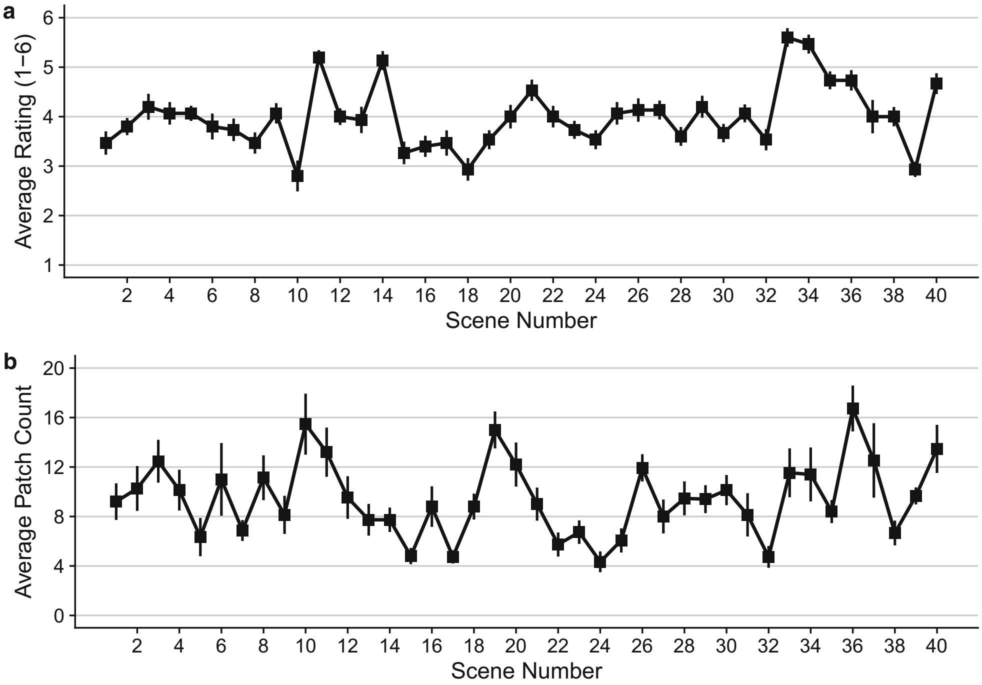


**Fig. 4** Response variability as a function of scene, showing average participant responses and standard errors of the responses as a function of scene for (a) the brightness rating task and (b) the patch count task

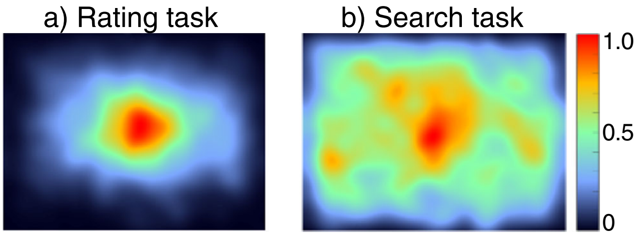


**Fig. 5** Center biased maps. Fixation density maps aggregated across participants and scenes are shown for (a) the brightness rating task and (b) the brightness search task

meaning accounted for 5% of the variance in fixation density after controlling for salience ($M$ = .05, $SD$ = .07), and salience accounted for 6% of the variance in fixation density after controlling for meaning ($M$ = .06, $SD$ = .07) (Fig. 7b). Again, this difference was not significant: $t$(78) = - 0.59, $p$ = .56, 95% CI = [- 0.04, 0.02]. Importantly, however, this pattern of results changed when using the unbiased meaning and saliency maps (Figs. 7c and d). Using the unbiased maps, meaning accounted for 22% of the overall variance in attention ($M$ = .22, $SD$ = .11), whereas salience explained only 4% of the variance ($M$ = .04, $SD$ = .05) among the linear correlations, $t$(78) = 6.42, $p$ < .001, 95% CI = [0.10, 0.18]. Similarly, for the semipartial correlations, meaning accounted for 18% of the total variance in attention ($M$ = .18, $SD$ = .11), whereas salience explained only 1% of the variance ($M$ = .04, $SD$ =

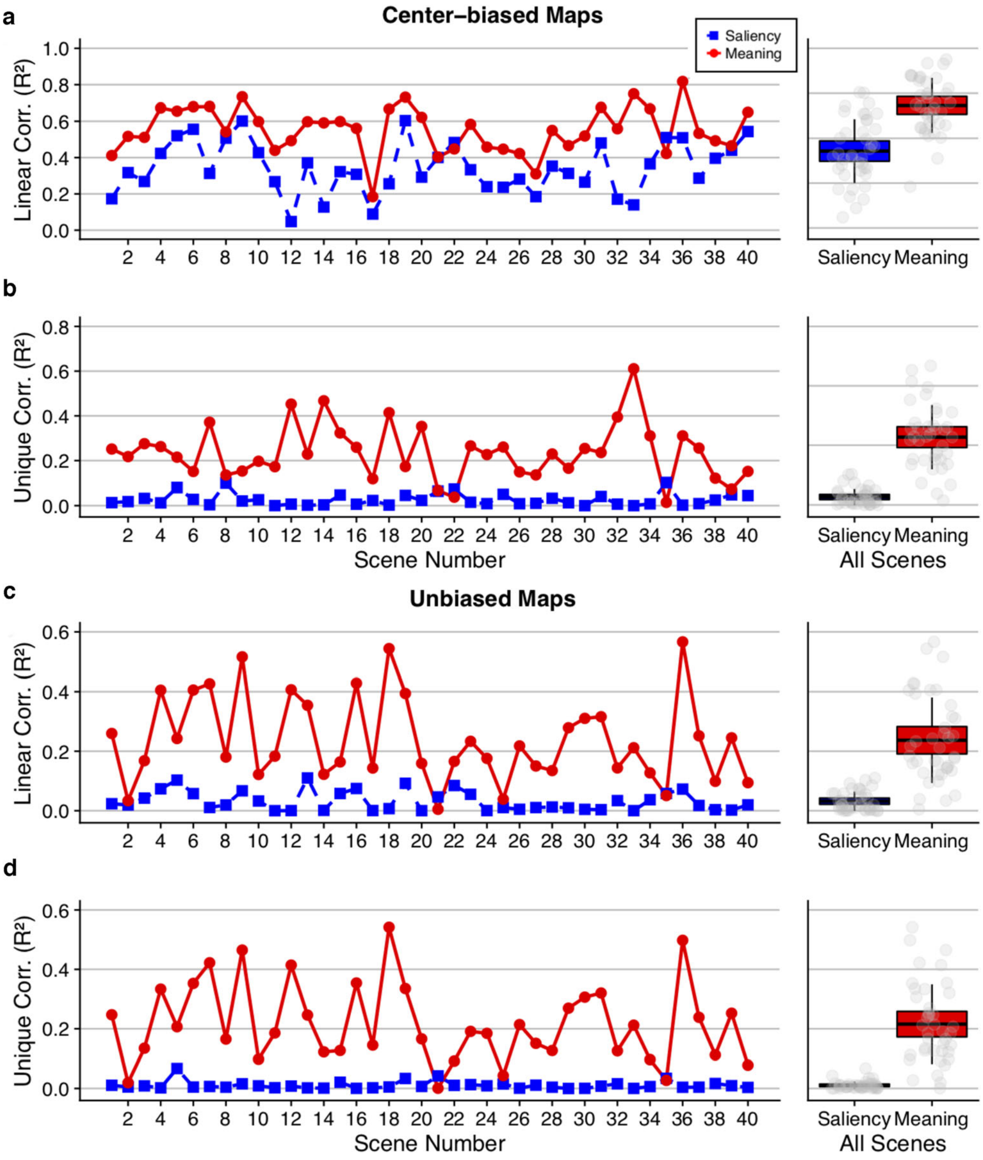


**Fig. 6** Squared linear correlations and semipartial correlations by scene for the brightness rating task. The upper line plots show the squared (a) linear and (b) semipartial correlations between the fixation density maps and both meaning (circles) and salience (squares) using center-biased meaning and saliency maps. The lower line plots show the squared (c) linear and (d) semipartial correlations using unbiased meaning and saliency maps. The scatterplots on the right show the grand mean (black horizontal lines), 95% confidence intervals (colored boxes), and one standard deviation (black vertical lines) for meaning and salience across all 40 scenes for each analysis

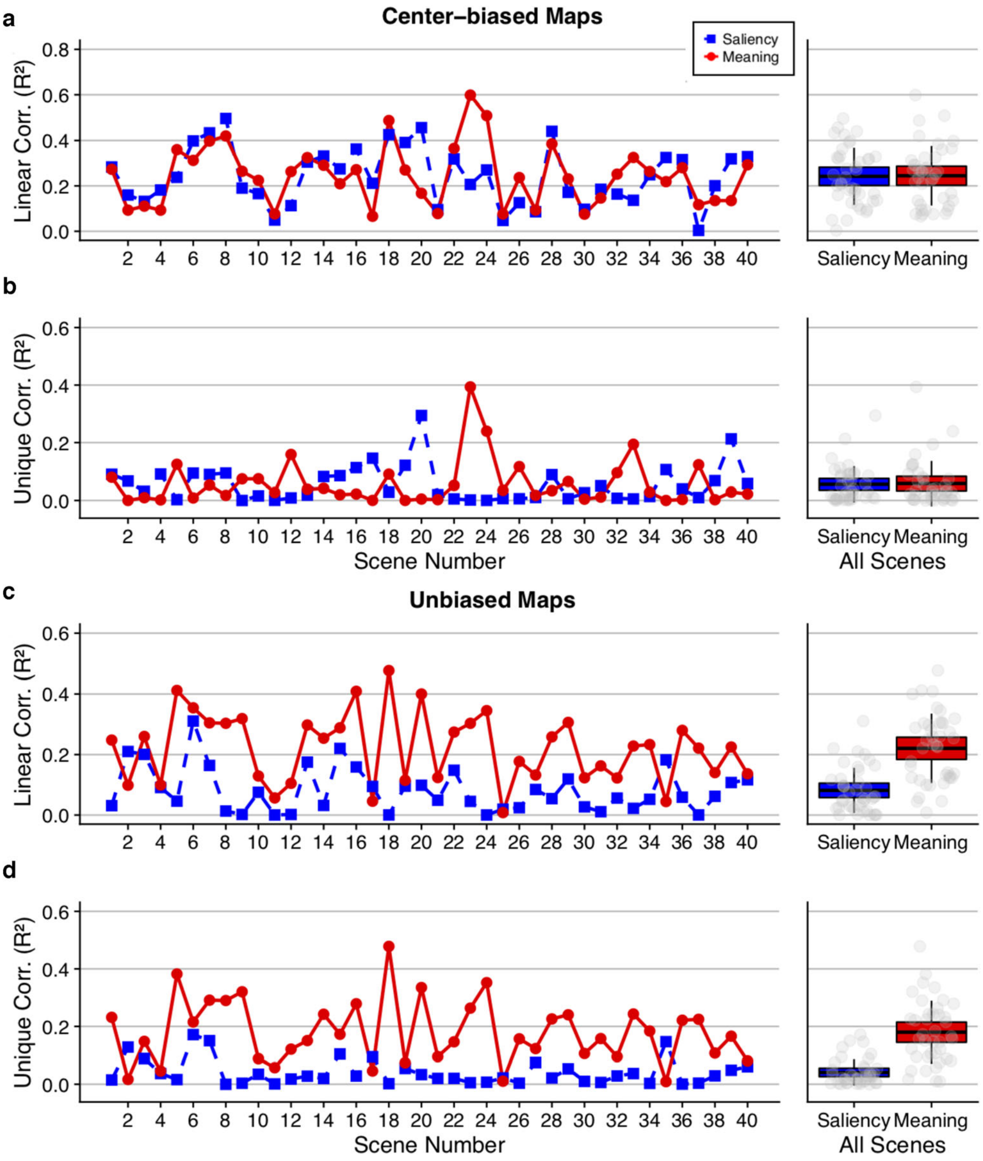


**Fig. 7** Squared linear correlations and semipartial correlations by scene for the brightness search task. The upper line plots show the squared (a) linear and (b) semipartial correlations between fixation density and both meaning (circles) and salience (squares) for the search task, using the center-biased meaning and saliency maps. The lower line plots show the squared (c) linear and (d) semipartial correlations for the search task using the unbiased meaning and saliency maps. The scatterplots on the right show the corresponding grand mean (black horizontal lines), 95% confidence intervals (colored boxes), and one standard deviation (black vertical lines) for meaning and salience across all scenes

.04), $t(78) = 7.42$, $p < .001$, 95% CI = [0.10, 0.18]. These findings suggest that when the more distributed nature of attention away from scene centers and to scene peripheries in the brightness search task was taken into account, meaning influenced attentional guidance more than salience, even though meaning was irrelevant and salience was central to the task.

## Fixation-by-fixation analyses

Previously, it has been posited that attention during scene viewing might initially be guided by salience, but that as time progresses, meaning begins to play an increasing role (Anderson, Donk, & Meeter, 2016; Anderson, Ort, Kruijne, Meeter, & Donk, 2015; Henderson & Ferreira, 2004; Henderson & Hollingworth, 1999; Parkhurst et al., 2002). On the other hand, in two studies investigating the roles of meaning and salience in memorization and scene description tasks, we did not observe this change from guidance by salience to guidance by meaning (Henderson & Hayes, 2017; Henderson et al., 2018). Instead, meaning was found to guide attention from the first saccade. Because the present tasks were designed to make meaning irrelevant and salience central, they provide another opportunity to test this hypothesis.

We conducted a temporal time-step analysis in which a series of attention maps were generated from each sequential fixation (first fixation, second fixation, third fixation, etc.) for each scene in each task. We then correlated each attention map for each fixation and scene using both the center-biased and unbiased meaning and saliency maps to calculate the squared linear and semipartial correlations. Then the correlations for each scene and fixation were averaged across scenes in order to assess how meaning and image salience predicted attention on a fixation-by-fixation basis. The prediction of the salience-first hypothesis is that the correlation between the saliency and attention maps should be greater for earlier than for later fixations, with salience dominating meaning in the earliest fixations.

**Brightness rating task** Using the center-biased maps, meaning accounted for 34%, 23%, and 17% of the variance in the first three fixations, whereas salience accounted for 8%, 12%, and 11% of the variance in the first three fixations, respectively, for the linear correlations (Fig. 8a). Two-sample, two-tailed *t* tests compared meaning and salience for all eight initial fixations, using *p* values corrected for multiple comparisons by using a false discovery rate (FDR) correction (Benjamini & Hochberg, 1995). Overall, this confirmed the advantage of meaning over salience for all eight fixations (all FDR *p*s < .05). Similarly, for the semipartial correlations, meaning accounted for 28%, 14%, and 9% of the variance in the first three fixations, and salience accounted for 2%, 3%, and 3% of the variance in the first three fixations (Fig. 8b). Again, meaning predicted attention significantly better than salience for all eight initial fixations (all FDR *p*s < .001). Using the unbiased maps, this overall pattern of results did not change (both linear and semipartial correlations: all FDR *p*s < .001) (Figs. 8c and 8d). These results do not support the hypothesis that the influence of meaning on attentional guidance was delayed until later fixations.

**Brightness search task** Using the center-biased maps, meaning accounted for 30%, 14%, and 7% of the variance in the first three fixations, and salience accounted for 11%, 16%, and 14% in the first three fixations, respectively, for the linear correlations (Fig. 9a). Here, meaning produced an advantage over salience on the first fixation (FDR *p* < .001), but not on Fixations 2 through 8 (FDR *p* > .05). For the semipartial correlations, meaning explained 22%, 8%, and 3% of the variance in the first three fixations, and salience accounted for 3%, 10%, and 10% in the first three fixations (Fig. 9b). Significant advantages were observed for meaning on the first fixation (FDR *p* < .001) and for salience on the third fixation (FDR *p* < .05), with no other comparisons reaching significance (FDR *p*s > .05).

Using the unbiased meaning and saliency maps, the pattern of results changed. For the linear correlations, meaning accounted for 5%, 6%, and 4% of the variance in the first three fixations, and salience accounted for 1%, 4%, and 5% of the variance in attention. Turning to the semipartial correlations, meaning accounted for 5%, 6%, and 3% of the variance in the first three fixations, and salience accounted for 0.1%, 3%, and 4% of the variance in attention. Meaning still produced an advantage over salience for the first fixation (linear and semipartial FDR *p*s < .05), with all other fixations showing nonsignificant differences (linear and semipartial FDR *p*s > .05). The advantage for salience over meaning for the third fixation seen in the center-biased maps was not observed with the unbiased maps.

The fixation-by-fixation analyses were not consistent with the salience-first hypothesis. In the analyses using both the center-biased and unbiased maps, meaning was more important than salience at the first fixation. Using the center-biased maps, salience was stronger at the third fixation. This result, however, was not true using the unbiased maps, suggesting that the advantage for salience in the center-biased maps was driven by the center bias rather than by salience itself. Overall, the results are not consistent with the hypothesis that attentional guidance transitions from salience to meaning over time.

## Saccade amplitude analyses

In the analyses thus far, fixations following both shorter and longer saccades were included. It could be that meaning guides attention within local scene regions, whereas salience guides attention as it moves from one scene region to another. To test this hypothesis, we analyzed the role of meaning in attentional guidance as a function of saccade amplitude. If meaning plays a greater role for local (e.g., within-object) shifts of attention, then meaning should be more related to attentional selection following shorter than following longer saccades. Such a pattern might be more likely to occur in the case of the present study, because meaning was not relevant to the tasks. To investigate this hypothesis, we assessed how both meaning and salience related to attention following saccades of shorter versus longer amplitudes (Fig. 10). Specifically, saccade amplitudes were binned by decile, and fixation density maps were created for each saccade amplitude decile. Meaning and saliency maps were then correlated with the fixation density maps for each decile. We conducted these analyses using both the center-biased and unbiased meaning and saliency maps. The saccade amplitude averages were 5.37° for the rating task (*SD* = 3.41) and 4.61° for the search task (*SD* = 3.51).

**Brightness rating task** For the brightness rating task, using the center-biased maps, meaning produced an advantage over salience for saccade amplitude Deciles 1 through 7 and 9 (FDR *p*s < .05), but not for Deciles 8 and 10 (FDR *p*s > .05). For the semipartial correlations, meaning explained significantly more of the variance in fixation density than did salience for all ten saccade amplitude deciles (all FDR *p*s < .05). When

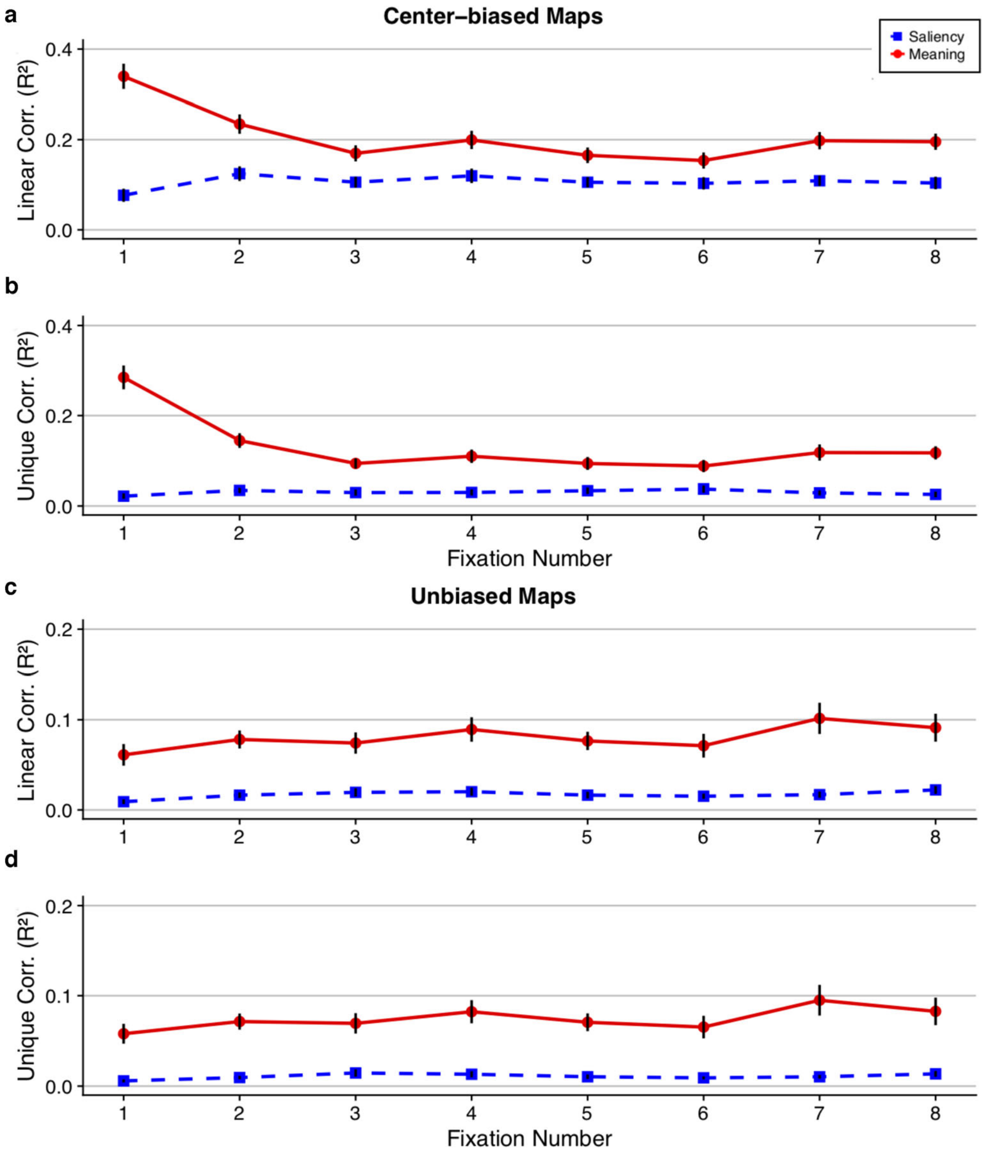


**Fig. 8** Fixation-by-fixation time-step analyses for the brightness rating task. The upper line plots show the squared (a) linear and (b) semipartial correlations between fixation density and both meaning (circles) and salience (squares) as a function of fixation number, collapsed across scenes for the rating task using the center-biased maps. The lower line plots show the squared (c) linear and (d) semipartial correlations between fixation density and both meaning (circles) and salience (squares) as a function of fixation order using the unbiased maps. Error bars represent the standard errors of the means

using the unbiased meaning and saliency maps, this pattern of results became stronger, as meaning produced an advantage over salience across all deciles in both the linear and semipartial correlations (FDR *p*s < .05).

**Brightness search task** For the brightness search task, using the center-biased maps, there were no significant differences between meaning and salience for any saccade amplitude decile in either the linear or the semipartial correlations (all FDR *p*s > .05). When using the unbiased maps, on the other hand, this pattern of results changed, as meaning produced an advantage over salience for saccade amplitude Deciles 1 through 9 (FDR *p*s < .05), but not for Decile 10 (FDR *p* > .05).

Overall, it appears that meaning was used to guide attention for both short and long shifts of attention, though there was some evidence that this influence was reduced when the scene peripheries were removed from the analyses (i.e., with the center-biased maps) and for the longest shifts of attention.

## Discussion

Past research has emphasized image salience as a key basis for attentional selection during real-world scene viewing (Borji et al., 2014; Borji et al., 2013; Harel et al., 2006; Itti & Koch, 2001; Koch & Ullman, 1985; Parkhurst et al., 2002).

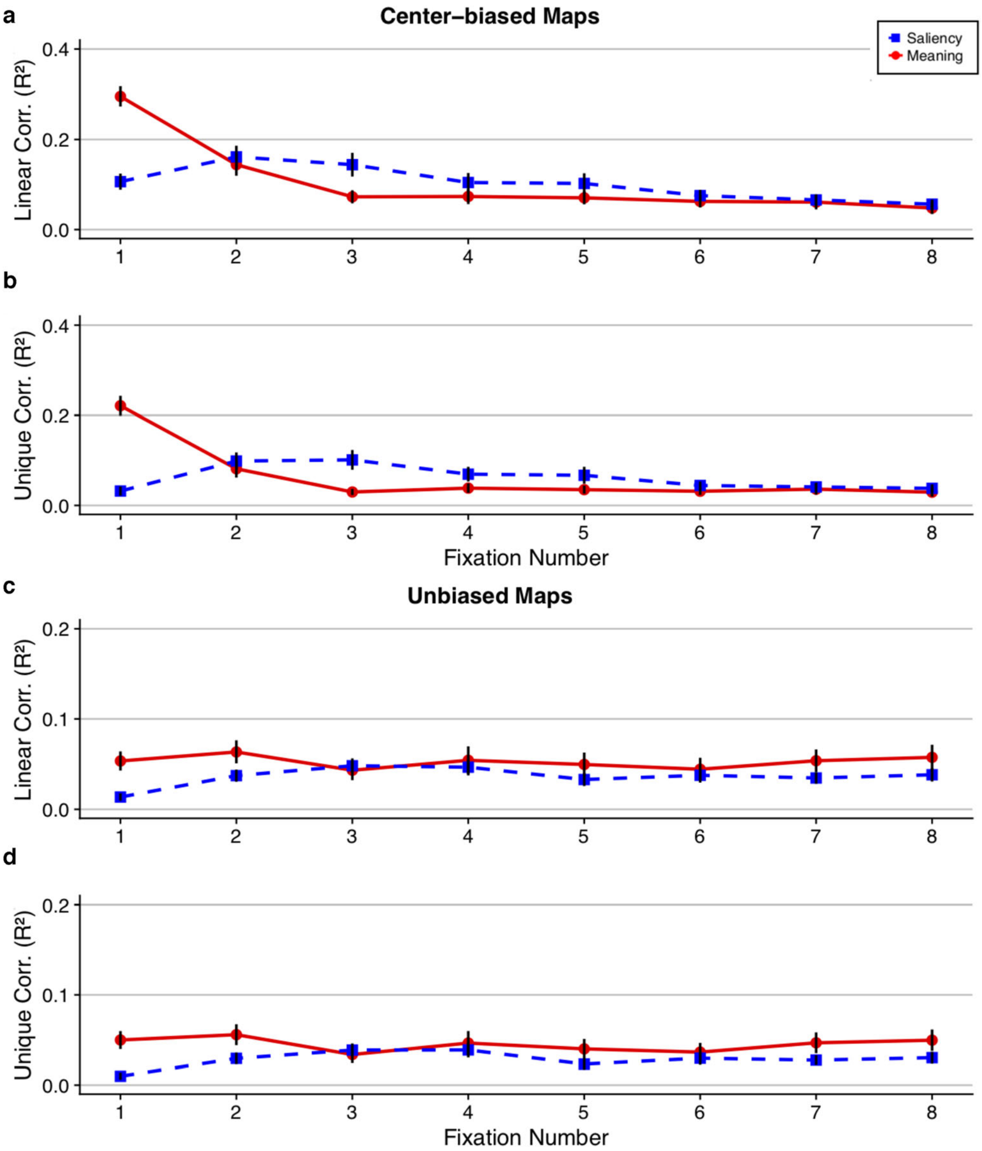
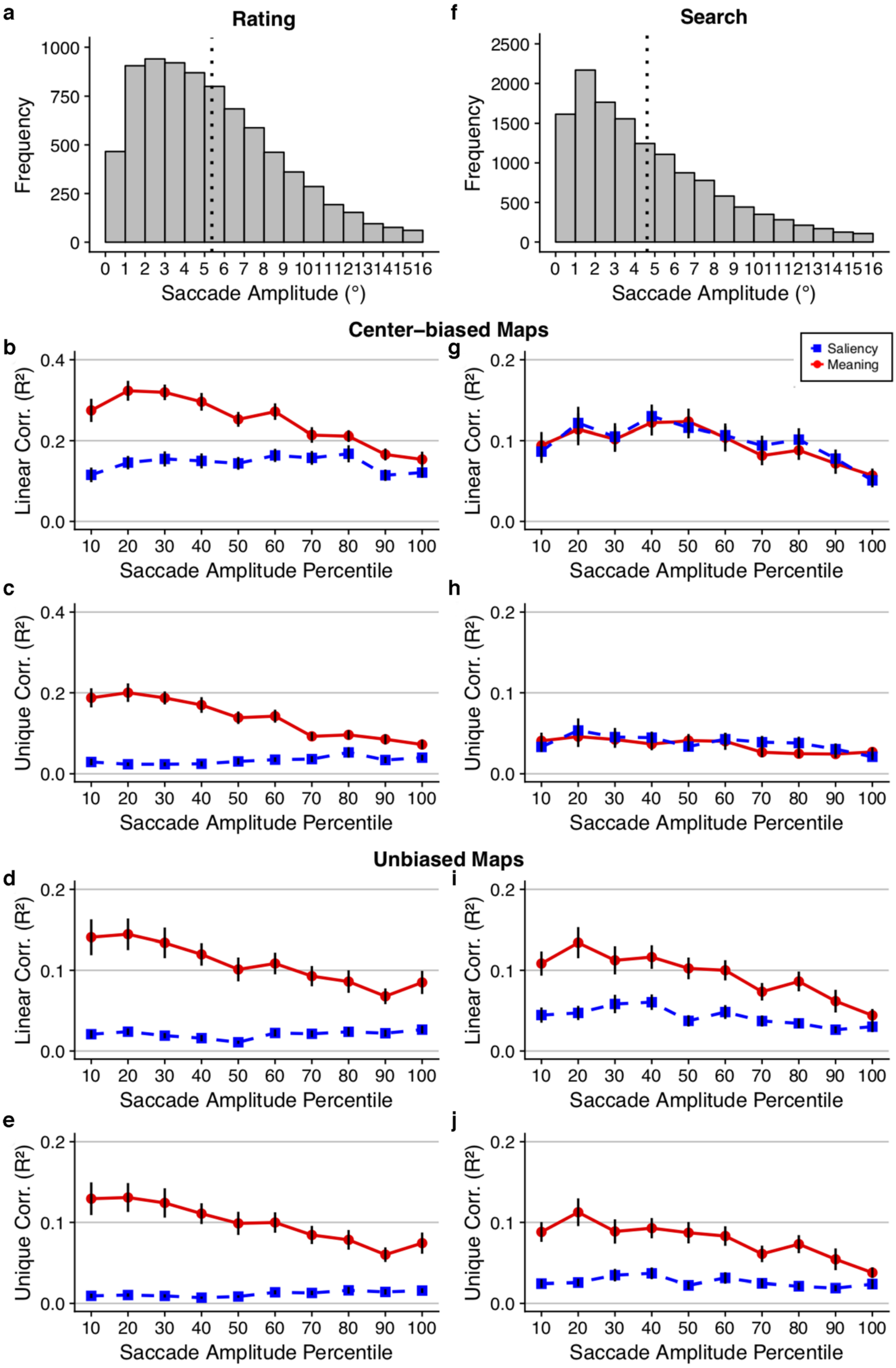


**Fig. 9** Fixation-by-fixation time-step analyses for the brightness search task. The upper line plots show the squared (a) linear and (b) semipartial correlations between fixation density and both meaning (circles) and salience (squares) as a function of fixation number, collapsed across scenes for the search task using the center-biased maps. The lower line plots show the squared (c) linear and (d) semipartial correlations between fixation density and both meaning (circles) and salience (squares) as a function of fixation order using the unbiased maps. Error bars represent the standard errors of the means

Although this previous work has provided an important framework for understanding attentional guidance in scenes, it downplays the fact that attention is strongly guided by cognitive factors related to the semantic features that are relevant to understanding the scene in the context of the task (Buswell, 1935; Hayhoe & Ballard, 2005; Hayhoe et al., 2003; Henderson et al., 2007; Henderson, Malcolm, & Schandl, 2009; Land & Hayhoe, 2001; Rothkopf et al., 2016; Yarbus, 1967). With the development of meaning maps, which capture the spatial distribution of semantic content in scenes in the same format in which saliency maps capture the spatial distribution of image salience, it has become possible to directly compare the influences of meaning and image salience on attention in scenes (Henderson & Hayes, 2017).

In prior studies comparing meaning and image salience during scene viewing, meaning has better explained the spatial and temporal patterns of attention (Henderson & Hayes, 2017, 2018; Henderson et al., 2018). However, those studies used memorization, aesthetic judgment, and scene description viewing tasks, and it could be argued that those tasks were biased toward attentional guidance by meaning. In the present study we sought to determine whether the influence of meaning on attention would be eliminated in tasks that do not require any semantic analysis of the scenes. To test this



hypothesis, we used two viewing tasks designed to eliminate the need for attending to meaning: a brightness rating task, in which participants rated the overall brightness of scenes, and a brightness search task, in which participants counted the number of bright areas in scenes.

For the brightness rating task, we found that meaning explained the spatial distribution of attention better than image salience. This result was observed both overall and when the correlation between meaning and image salience was statistically controlled, and it held for early scene viewing, for short and long saccades, and when using center-biased and unbiased meaning and saliency maps. For the brightness search task using center-biased meaning and saliency maps, there were no differences between meaning and salience, either

◀ **Fig. 10** Squared linear correlations and squared semipartial correlations as a function of saccade amplitude to a fixation. The saccade amplitude results for the rating task are shown in the first column (panels a through e), in which panel a shows a histogram of saccade amplitude frequencies and average saccade amplitude (black dotted line). Panels b and d show the squared linear correlations, and panels c and e the semipartial correlations, between both meaning (circles) and salience (squares) and fixation density, as a function of saccade amplitude percentiles prior to fixation for the center-biased maps (panels b and c) and the unbiased maps (panels d and e). The second column (panels f through j) shows the saccade amplitude results for the search condition, in which panel f shows a histogram of saccade amplitude frequencies and average saccade amplitude (black dotted line). Panels g and i show the squared linear correlations, and panels h and j the semipartial correlations, between both meaning (circles) and salience (squares) and fixation density, as a function of saccade amplitude percentiles using the center-biased maps (panels g and h) and the unbiased maps (panels i and j). The data points are averaged across all 40 scenes at each decile. Error bars represent standard errors of the means

overall or when controlling for their correlation. However, the center-biased maps did not capture the fact that during the search task, the center bias in attention was greatly attenuated, because attention was distributed much more uniformly over the scenes. Meaning and saliency maps with center bias overweight scene centers and ignore scene peripheries, which was opposite to the attention maps we actually observed. When the attention maps were analyzed using meaning and saliency maps that did not include center bias, the results were similar to those of the brightness rating task: Meaning explained the variance in attention better than salience, both overall and after statistically controlling for the correlation between meaning and salience. This pattern held for both short and long saccades and for the first saccade.

Overall, the results provide strong evidence that the meaning of a scene plays an important role in guiding attention through real-world scenes, even when meaning is irrelevant and image salience is relevant to the task. Converging evidence across two viewing tasks that focused on an image property related to image salience showed that meaning accounted for more variance in attentional guidance than did salience and, critically, that when the correlation between meaning and salience was controlled, only meaning accounted for significant unique variance. These results indicate that the guidance of attention by meaning is not restricted to viewing tasks that focus on encoding the meaning of the scene, strongly suggesting a fundamental role of meaning in attentional guidance within scenes.

Although the main pattern of results was clear and generally consistent across the two tasks, a few points are worth additional comment. First, our results suggest that tasks can differ in the degree to which a center bias is present. Here, center bias was much greater when judging overall scene brightness than when searching for bright scene regions. These differences in center bias for the rating and search tasks likely occurred due to differences in the requirements of the tasks. The rating task simply required participants to rate the overall brightness of scenes, so there was no particular reason for viewers to direct attention away from the centers and to the peripheries of the scenes. In comparison, the search task required participants to count individual bright regions, many of which appeared away from the scene centers and in the peripheries. This resulted in fewer central fixations and more peripheral fixations in the brightness search task than in the brightness rating task. Because there were more peripheral fixations in the search task, the center-biased meaning and saliency maps did not have the same predictive power to capture the relationship between meaning, salience, and attention as they did for the brightness rating task. Indeed, for this reason, neither meaning nor saliency maps did a particularly good job of predicting attention when center bias was included in the maps. However, when the center bias was removed from the two prediction maps, meaning maps were significantly better than saliency maps in accounting for attention.

The difference between the center-biased and unbiased maps was also evident in the analysis focusing on the earliest eye movements. According to the "salience-first" hypothesis, we should have seen an initial bias of attention toward salient regions, followed by a shift to meaningful regions. In our prior studies, we instead observed that meaning guided attention from the very first eye movement (Henderson & Hayes, 2018; Henderson et al., 2018). In the present study, when center bias was included in the meaning and saliency maps in the brightness search task, meaning initially guided attention in the first eye movement, but there was a tendency for salience to take over for a few saccades before meaning again dominated. This pattern might offer some small support for salience first. However, as we noted, viewers were much less likely to attend to scene centers and more likely to move their eyes to the edges of the scenes in the brightness search task. When the unbiased maps were used in the search task analysis, the trend from meaning to salience over the first few fixations was not observed. At best, then, there is a hint that when the viewer's task is explicitly to find and count salient scene regions, viewers may be slightly more biased early on to attend to regions that are more salient. However, this result is weak at best, given that it appeared only in the third fixation and disappeared in the unbiased map analysis. Overall, even in a task that explicitly focused on salience and in which meaning was completely irrelevant, meaning played a stronger role in attentional guidance from the very beginning of viewing.

The type of meaning studied in the present work is what we refer to as *context-free* meaning, in that it is based on ratings of the recognizability and informativeness of isolated scene patches, shown to raters independently of the scenes from which they are derived and independently of any task or goal besides the rating itself. Other types of meaning may be of interest in future studies. For example, we can consider *contextualized* meaning, in which meaning is determined on the basis of how important a scene patch is with respect to its global scene context. Additionally, the role of task may affect meaning, as well. For example, meaning within a scene may change depending on a viewer's current tasks or goals. Because meaning can be defined in so many ways, it is necessary that we understand how these variants influence attentional guidance. The meaning map approach provides a method for pursuing these important questions.

## Conclusion

We investigated the relative importance of meaning and image salience in attentional guidance within scenes, using tasks that do not require semantic analysis and in which salience plays a critical role. Overall, the results strongly suggested that viewers can't help but attend to meaning (Greene & Fei-Fei, 2014). These findings are most consistent with cognitive control theories of scene viewing, in which attentional priority is assigned to scene regions on the basis of semantic properties rather than image properties.

**Author note** We thank Conner Daniels, Praveena Singh, and Diego Marquez for assisting in data collection. This research was supported by the National Eye Institute of the National Institutes of Health, under award number R01EY027792. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors declare no competing financial interests. Additionally, these data have not previously been published.

## References

Anderson, N. C., Donk, M., & Meeter, M. (2016). The influence of a scene preview on eye movement behavior in natural scenes. *Psychonomic Bulletin & Review*, *23*, 1794–1801. https://doi.org/10.3758/s13423-016-1035-4

Anderson, N. C., Ort, E., Kruijne, W., Meeter, M., & Donk, M. (2015). It depends on when you look at it: Salience influences eye movements in natural scene viewing and search early in time. *Journal of Vision*, *15*(5), 9:1–22. https://doi.org/10.1167/15.5.9

Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology*, *103*, 62–70. https://doi.org/10.1037/h0036799

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, *57*, 289–300.

Borji, A., Parks, D., & Itti, L. (2014). Complementary effects of gaze direction and early saliency in guiding fixaitons during free viewing. *Journal of Vision*, *14*(13), 3. https://doi.org/10.1167/14.13.3

Borji, A., Sihite, D. N., & Itti, L. (2013). Quantative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, *22*, 55–69. https://doi.org/10.1109/TIP.2012.2210727

Buswell, G. T. (1935). How people look at pictures: A study of the psychology of perception in art. Chicago: University of Chicago Press.

Carmi, R., & Itti, L. (2006). The role of memory in guiding attention during natural vision. *Journal of Vision*, *6*(9), 4:898–914. https://doi.org/10.1167/6.9.4

Greene, M. R., & Fei-Fei, L. (2014). Visual categorization is automatic and obligatory: Evidence from Stroop-like paradigm. *Journal of Vision*, *14*(1), 14. https://doi.org/10.1167/14.1.14

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.), Advances in neural information processing systems: Proceedings of NIPS'06 (pp. 545–552). Cambridge, US: MIT Press.

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

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, *3*(1), 6:49–63. https://doi.org/10.1167/3.1.6

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

Henderson, J. M. (2017). Gaze control as prediction. *Trends in Cognitive Sciences*, *21*, 15–23. https://doi.org/10.1016/j.tics.2016.11.003

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. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), Eye movements: A window on mind and brain (pp. 537–562). Amsterdam: Elsevier.

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. M. Henderson & F. Ferreira (Eds.), The interface of language, vision, and action: Eye movements and the visual world (pp. 1–58). New York: Psychology Press.

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

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision*, *18*(6), 10:1–18. https://doi.org/10.1167/18.6.10

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

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, *50*, 243–271. https://doi.org/10.1146/annurev.psych.50.1.243

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. https://doi.org/10.3758/PBR.16.5.850

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

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*, 1254–1259. https://doi.org/10.1109/34.730558

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

Land, M. F., & Hayhoe, M. (2001). In what ways to eye movements contribute to everyday activities? *Vision Research*, *41*, 3559–3565. https://doi.org/10.1016/S0042-6989(01)00102-X

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics*, *2*, 547–552. https://doi.org/10.3758/BF03210264

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, *45*, 205–231. https://doi.org/10.1016/j.visres.2004.07.042

Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron*, *53*, 605–617. https://doi.org/10.1016/j.neuron.2007.01.018

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

Rahman, S., & Bruce, N. (2015). Visual saliency prediction and evaluation across different perceptual tasks. *PLoS ONE*, *15*, e138053. https://doi.org/10.1371/journal.pone.0138053

Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2016). Task and context determine where you look. *Journal of Vision*, *7*(14), 16:1–20. https://doi.org/10.1167/7.14.16

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5), 5. https://doi.org/10.1167/11.5.5

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*, 766–786. https://doi.org/10.1037/0033-295X.113.4.766

Yarbus, A. L. (1967). Eye movements during perception of complex objects. In Eye movements and vision (pp. 171–211). New York: Plenum Press.