Overt attentional correlates of memorability of scene images and their relationships to scene semantics

Muxuan Lyu* Department of Management and Marketing, The Hong Kong Polytechnic University, Hong Kong, China

Kyoung Whan Choe* Department of Psychology, The University of Chicago, Chicago, IL, USA; Mansueto Institute for Urban Innovation, The University of Chicago, Chicago, IL, USA

Omid Kardan Department of Psychology, The University of Chicago, Chicago, IL, USA

Hiroki P. Kotabe Kotabe Labs, Chicago, IL, USA

John M. Henderson Center for Mind and Brain and Department of Psychology, University of California, Davis, Davis, CA, USA

Marc G. Berman Department of Psychology, The University of Chicago, Chicago, IL, USA; Grossman Institute for Neuroscience, Quantitative Biology and Human Behavior, The University of Chicago, Chicago, IL, USA

Computer vision-based research has shown that scene semantics (e.g., presence of meaningful objects in a scene) can predict memorability of scene images. Here, we investigated whether and to what extent overt attentional correlates, such as fixation map consistency (also called inter-observer congruency of fixation maps) and fixation counts, mediate the relationship between scene semantics and scene memorability. First, we confirmed that the higher the fixation map consistency of a scene, the higher its memorability. Moreover, both fixation map consistency and its correlation to scene memorability were the highest in the first 2 seconds of viewing, suggesting that meaningful scene features that contribute to producing more consistent fixation maps early in viewing, such as faces and humans, may also be important for scene encoding. Second, we found that the relationship between scene semantics and scene memorability was partially (but not fully) mediated by fixation map consistency and fixation counts, separately as well as together. Third, we found that fixation map consistency, fixation counts, and scene semantics significantly and additively contributed to scene memorability. Together, these results suggest that eye-tracking measurements can complement computer vision-based algorithms and improve overall scene memorability prediction.

Introduction

Some visual scenes are more memorable than other scenes (Isola, Xiao, Torralba, & Oliva, 2011). Investigating scene memorability not only is important for understanding human vision and memory but is also useful for people interested in predicting and maximizing it for practical purposes. Computer vision-based research (Isola, Xiao, Torralba, & Oliva et al., 2011; Khosla, Raju, Torralba, & Oliva, 2015) has shown that intrinsic features of a scene, such as its semantics, global descriptors, object counts or areas, interestingness, and aesthetics, can affect scene memorability in a similar manner across different viewers. However, scene memory can also be modulated by factors extrinsic to a scene, such as the other scenes that were presented with it (Bylinskii, Isola, Bainbridge,

Citation: Lyu, M., Choe, K. W., Kardan, O., Kotabe, H. P., Henderson, J. M., & Berman, M. G. (2020). Overt attentional correlates of memorability of scene images and their relationships to scene semantics. *Journal of Vision*, *20*(9):2, 1–17, https://doi.org/10.1167/jov.20.9.2.

https://doi.org/10.1167/jov.20.9.2 Received July 19, 2019; published September 2, 2020 ISSN 1534-7362

Torralba, & Oliva, 2015) and the viewing tasks that were performed while each scene was presented (Choe, Kardan, Kotabe, Henderson, & Berman, 2017; Wolfe, Horowitz, & Michod, 2007). Despite great research interest, factors that could contribute to scene memorability are not fully understood.

Eye tracking enables the investigation of underlying attentional mechanisms of scene memory. Particularly, fixation count has been repeatedly demonstrated to be associated with scene memory. For example, an increased fixation count during encoding is associated with better recognition on a trial-by-trial basis for scenes (Choe et al., 2017) and objects (Tatler & Tatler, 2013), suggesting that trial-level fixation counts signals viewers' elaborate inspection of a scene and that elaborate inspection can enhance scene encoding (Winograd, 1981). Moreover, more preferred scenes produce more fixations and are better remembered later than less preferred scenes (Loftus, 1972), suggesting that population-level fixation counts (i.e., the averaged fixation counts across viewers) can be a proxy for an intrinsic property of a scene, such as interestingness (e.g., the more interesting a scene is, the more elaborate inspection viewers do). Together, these results suggest that fixation count is still a very important source of information for studying the attentional mechanisms of scene encoding, despite its simplicity.

Eye tracking also enables the investigation of the relationship between where viewers look in scenes (i.e., fixation maps) (Henderson, 2003; Pomplun, Ritter, & Velichkovsky, 1996; Wooding, 2002) and how scene memory is formed (Choe et al., 2017; Hollingworth, 2012; Olejarczyk, Luke, & Henderson, 2014; Ramey, Henderson, & Yonelinas, 2020; Tatler & Tatler, 2013). For example, the fixation map from a scene during intentional memorization is different from that during visual search (Castelhano, Mack, & Henderson, 2009), and the degree of difference in the fixation maps during memorization versus visual search in the same scene could explain how visual search impaired incidental scene memory on a trial-by-trial basis (Choe et al., 2017). Similar to these approaches, one can also examine the consistency of fixation maps across viewers, also called inter-observer congruency or inter-subject consistency (i.e., fixation map consistency) (Dorr, Martinetz, Gegenfurtner, & Barth, 2010; Torralba, Oliva, Castelhano, & Henderson, 2006), which is a scene-specific, population-level measure (i.e., averaged over a group of participants for each scene) and often used in evaluating computational fixation prediction models by providing an upper bound of the performance that those models can achieve (Wilming, Betz, Kietzmann, & König, 2011). Importantly, two previous papers briefly reported that it is positively associated with scene memorability (Khosla et al., 2015; Mancas & Le Meur, 2013).

In this study, we used two different eye-tracking datasets, the Edinburgh dataset (Luke, Smith, Schmidt, & Henderson, 2014; Nuthmann & Henderson, 2010; Pajak & Nuthmann, 2013) and the FIGRIM dataset (Bylinskii et al., 2015) to investigate the relationships among fixation map consistency, fixation counts, scene semantics, and scene memorability. Both datasets included eye-tracking data from a group of participants engaged in scene encoding, and the measures of scene memorability came from a different group of participants engaged in scene recognition tasks (i.e., averaged recognition accuracy across participants in these tasks). An advantage of the Edinburgh dataset was the 8 seconds of scene viewing duration (vs. 2 seconds in the FIGRIM dataset). An advantage of the FIGRIM dataset was its extensive object annotations, which can be used to obtain proxies of scene semantics (Isola et al., 2011; Xu, Jiang, Wang, Kankanhalli, & Zhao, 2014). We exploited both commonalities and unique strengths of these two datasets to ask the following questions.

First, using both the Edinburgh and FIGRIM datasets, we examined whether and to what extent fixation map consistency and fixation counts, respectively, are associated with scene memorability. Second, using the FIGRIM dataset, we asked whether and to what extent fixation map consistency and fixation counts mediate the relationship between scene semantics and scene memorability, after confirming that scene semantics and scene memorability were associated as expected. Third, using the FIGRIM dataset, we tested whether there were additive and/or interactive effects of fixation map consistency, fixation count, and scene semantics on scene memorability. Fourth, using the Edinburgh dataset, we examined the effects of viewing time on fixation map consistency (Buswell, 1935; Tatler, Baddeley, & Gilchrist, 2005)—that is, the temporal consistency of fixation maps across participants and within the same participant and how fixation maps may be related to scene memorability. Finally, we quantified center bias in both the Edinburgh and the FIGRIM scenes (Bindemann, 2010; Hayes & Henderson, 2020; Tatler, 2007; Tseng, Carmi, Cameron, Munoz, & Itti, 2009) using low-level visual saliency (Harel, Koch, & Perona, 2007) and examined its effect on fixation map consistency and scene memorability.

Consistent with previous studies (Khosla et al., 2015; Mancas & Le Meur, 2013), we confirmed a robust relationship between fixation map consistency and scene memorability in both datasets. Importantly, we also found that the relationships between scene semantics and scene memorability were partially (but not fully) mediated by fixation map consistency and fixation counts, separately as well as together. Additionally, we found that fixation map consistency, fixation counts, and scene semantics additively contribute to scene memorability where each contributes unique variance

in explaining scene memorability. These results suggest that eye-movement data add signal beyond scene semantics in the prediction of scene memorability.

Methods

Overview

This study is a re-analysis of two previously collected eye-tracking datasets, the sample sizes of which were determined for different purposes. All of our data and analysis codes are available at https://osf.io/hvgk6/.

The Edinburgh dataset (Luke et al., 2014; Nuthmann & Henderson, 2010; Pajak & Nuthmann, 2013) has been used in prior publications (Choe et al., 2017; Einhäuser & Nuthmann, 2016; Kardan, Berman, Yourganov, Schmidt, & Henderson, 2015; Kardan, Henderson, Yourganov, & Berman, 2016; Nuthmann, 2017) and is available from the author J.M.H. upon request. This dataset has the fixation map patterns of 135 scenes under three different encoding tasks (intentional memorization, visual search, and aesthetic preference evaluation) from 72 participants and memorability scores of these scenes from a subset of the participants (36). Out of 135 scenes, we analyzed only the 132 scenes that were used in both the encoding tasks and memory test (the scenes are available at https://osf.io/hvgk6/). In addition, we analyzed only the fixation data during the intentional memorization task from the 24 participants who performed this task on the 132 scenes, resulting in 24 fixation maps per scene and recognition accuracy from 12 participants per scene. For experimental details, please see Supplementary Note S1.

Second, the FIGRIM dataset (Bylinskii et al., 2015), which is freely available at https://github.com/cvzoya/figrim, has eye movement data from 67 in-lab participants viewing 630 scenes across 21 different categories (30 scenes per category) and memorability scores from 74 Amazon Mechanical Turk (AMT) participants. For experimental details, please see Bylinkskii et al. (2015).

Scene memory tasks and scene memorability definition

The Edinburgh study used a surprise scene memory test after completing all three scene encoding task blocks (intentional memorization, visual search, and preference evaluation). Before the task, participants were informed that their memory would be tested for all of the scenes that they had previously encountered, not just the scenes they had been instructed to remember in the memorization block. In each trial of the memory test, a scene was shown for 3 seconds, and participants were asked to identify whether the scene was "old" (encountered in the encoding phase during any block, not just the memorization block, and presented in an identical form), "altered" (encountered in the encoding phase but presented in a horizontally mirrored form), or "new." Whether or not a scene was horizontally flipped in the recognition test (i.e., scene orientation) was found to affect recognition accuracy (Choe et al., 2017), so it was included in our analyses. In the 132 scenes we analyzed, 66 scenes were "old," 66 scenes were "altered," and none was "new." Scene memorability was calculated as the average recognition accuracy in the memory test: the number of hit (correctly recognized) trials divided by the number of hit and miss trials, which equaled the number of participants (12) who saw these scenes.

The FIGRIM study used a continuous scene recognition task (Isola et al., 2011), in which participants were shown a series of new and repeated scenes and asked to press a key whenever they recognized a repeat scene. Scene memorability was calculated as the number of hit trials (trials where participants correctly pressed a button to a repeat scene) divided by the sum of both hit trials and miss trials (trials where participants did not press a button to a repeat scene) across participants.

Eye movement analysis

Edinburgh dataset

The raw eye movement data, sampled at 1000 Hz, were preprocessed using Eyelink Data Viewer (SR Research, Kanata, Canada) to identify discrete fixations and fixation durations during 8 seconds of scene viewing. Fixations were excluded from analysis if they were preceded by or co-occurred with blinks, were the first or last fixation in a trial, or had durations less than 50 ms or longer than 1200 ms. Fixation counts are the number of discrete fixations, regardless of their duration, that landed on the scenes.

FIGRIM dataset

Bylinkskii et al. (2015) "processed the raw eye movement data using standard settings of the EyeLink Data Viewer to obtain discrete fixations, removed all fixations shorter than 100 ms or longer than 1500 ms, and kept all others that occurred within the 2000 ms recording segment (from image onset to image offset)." The sampling rate was 500 Hz. The openly available FIGRIM dataset does not contain fixation duration information. Fixation counts are the number of discrete fixations on the scenes.

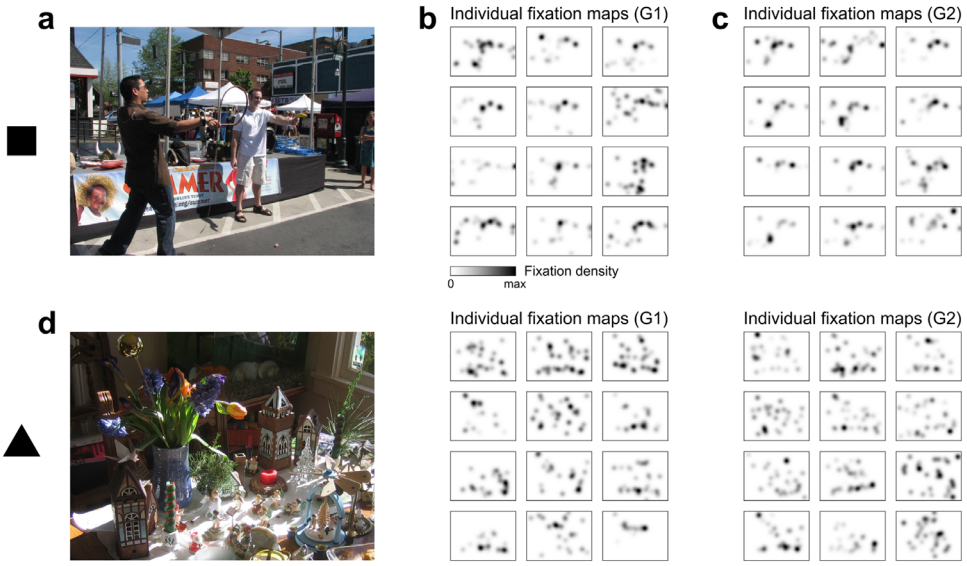

Figure 1. Individual fixation maps of two example scenes during intentional memorization from the Edinburgh dataset. (a) A scene that produced highly consistent fixation maps. The filled square on the left is used in Figure 3a to indicate this scene. (b) Fixation maps of 12 G1 participants, who were asked to memorize the scene and tested for their scene memory in the following recognition test (see Methods). The average recognition accuracy of these participants was used as scene memorability. (c) Fixation maps of 12 G2 participants, who were asked to memorize the scene but were not tested for their memory. (d) Individual fixation maps from G1 (middle panel) and G2 (right panel) for another scene (left panel) that produced less consistent fixation maps. The triangle on the left is in Figure 3a to indicate this scene.

Fixation map analysis

We used custom MATLAB (MathWorks, Natick, MA) scripts to do following. An individual fixation map of a participant viewing a scene (Figure 1) was constructed by convolving a Gaussian kernel over its duration-weighted fixation locations during 8 seconds of viewing (the Edinburgh dataset) or over equal-weighted fixation locations during 2 seconds of viewing (the FIGRIM dataset). The full width at half maximum of the Gaussian kernel was set to 2° (i.e., $\sigma = 0.85°$) to simulate central foveal vision and to take into account the measurement errors of video-based eye trackers (Choe, Blake, & Lee, 2016; Wyatt, 2010).

Fixation map consistency

The similarity of individual fixation maps across multiple viewers was quantified as in previous research (Dorr et al., 2010; Torralba et al., 2006). For each individual fixation map, its similarity to the averaged fixation map of the other leave-one-out fixation maps was calculated; then, the similarity values of all fixation maps were averaged to yield fixation map consistency. For example, in Figure 1b, 12 similarity values were obtained by comparing each individual fixation map versus the average of the other 11 fixation maps; those 12 values were then averaged to produce a fixation map consistency score. For the similarity metric, we opted for the Fisher z-transformed Pearson's correlation coefficient (Choe et al., 2017), among several metrics on fixation and saliency maps (see Dorr et al., 2010, and Le Meur & Baccino, 2013), because it is invariant to linear transformations, such as scaling.

Multivariate object presence score analysis

We relied on the extensive object annotations in the FIGRIM dataset for this analysis. Among the 707 objects that were included in the dataset, we selected 98 objects that appeared in at least 10 (out of 630) scenes, as Isola and colleagues (2011) did. We checked whether each object was present in each scene and coded its presence as 1 and absence as 0 for that scene. Then, we used the presence of these 98 annotated objects to predict scene memorability (i.e., AMT recognition accuracy) with a leave-one-out multivariate regression analysis. Specifically, for each test scene, we left out the object presence information of that scene, used the information of the remaining 629 scenes to train a regression model for predicting scene memorability, and used the leave-one-out model to predict scene memorability of the test scene, dubbed the multivariate object presence score (MOPS). We obtained MOPS for all of the 630 scenes by repeating the leave-one-out analysis. The correlation between

the MOPS and AMT recognition accuracy was 0.37 (95% CI, 0.30–0.44; $p < 0.001$).

To examine whether the relationship between MOPS and scene memorability was mediated by fixation map consistency and fixation count, we conducted a parallel mediation analysis using the *mediate* function with 5000 bootstrap resampling in the Psych Package (Revelle, 2020) in R (R Core Team, 2017). Scene memorability, MOPS, fixation map consistency, and fixation counts were all *z*-scored before conducting the mediation analysis. The mediation effect occurs when the indirect effect is significant (i.e., its confidence interval does not include zero). Full mediation occurs when the direct effect of the predictor variable is no longer significant ($p > 0.05$) by introducing the mediating variable(s), and partial mediation occurs when the direct effect of the predictor variable is still significant ($p < 0.05$) but significantly weakened by introducing the mediating variable(s).

Individual object presence analysis

In addition to the presences of 98 objects appearing in at least 10 FIGRIM scenes, we manually coded three scene semantic features: the presences of face/human, motion, and watchability in each scene (i.e., 1 if present, 0 if absent), following Xu et al. (2014). Specifically, the presence of face/human was based on whether a scene had humans or the faces of humans, animals, or objects that have facial features in a coherent manner like a giant face on the building or Thomas the train. One hundred fifty-eight FIGRIM scenes included face/human. The presence of motion was based on whether a scene contained moving or flying objects including humans or animals with meaningful gestures. One hundred nineteen FIGRIM scenes included motion, and 70 out of the 119 scenes also included face/human. The presence of watchability was based on whether a scene contained man-made objects designed to be watched (e.g., a display screen). Two hundred seventy-six FIGRIM scenes included watchability. We next examined whether or not the presence of each object/feature affected scene memorability significantly after correcting for multiple comparisons (correcting for 101 tests; 98 objects and three manually coded face/human, motion, and watchability) by using the unequal variance *t*-test and Holm–Bonferroni procedure. Supplementary Table S1 shows the top 20 objects/features in the descending order of mean difference (*t*-score) in scene memorability. The presences of face/human, person, pilot, and motion significantly increased scene memorability after correcting for multiple comparisons. Because the presence of face/human had the most scenes, mostly overlapped with person, pilot, and motion (91 out of 101 scenes with person, 10 out of 10 pilot scenes, and 70 out of 119 scenes with motion were also coded as face/human) and explained away the effect of motion on scene memorability, we examined only the presence of face/human. After seeing that it significantly increased both fixation map consistency and fixation count, we conducted a parallel mediation analysis to examine whether and to what extent these eye-tracking measures could mediate the relationship between the presence of face/human and scene memorability.

Center bias analysis

We calculated center bias for each scene, following Hayes and Henderson (2020). First, saliency maps for each scene (Figure 2b) were computed using the Graph-Based Visual Saliency (GBVS) algorithm (Harel, Koch, & Perona, 2007) with default settings and then normalized to make the total sum of all pixel values equal the number of pixels (800×600 pixels for the Edinburgh scenes and 1000×1000 pixels for the FIGRIM). The MATLAB code for the GBVS algorithm was obtained from http://www.vision.caltech.edu/~harel/share/gbvs.php. Next, an element-wise multiplication was performed between the saliency map and a Gaussian kernel (Figure 2c) with the σ of 10% of the scene height (i.e., 60 pixels for the Edinburgh scenes and 100 pixels for the FIGRIM scenes), which was done to downweight the saliency scores in the periphery pixels to account for the central fixation bias (Bindemann, 2010; Hayes & Henderson, 2020; Tatler, 2007; Tseng et al., 2009). The center bias of each scene (the white numbers in the right bottom in Figure 2d) was calculated by adding up the downweighted scores of all its pixels, which could differentiate the scenes with strong center bias from those with weak center bias. Figures 2e and 2f show the distributions of center bias in the Edinburgh and FIGRIM datasets, respectively.

Statistical software

The fixation map and center bias analyses were performed using MATLAB R2015b and custom MATLAB scripts, available at https://osf.io/hvgk6/. The *fitlm* function was used to perform linear regression analyses, the *anova1* function was used to perform one-way analysis of variance (ANOVA), and the *corrcoef* function was used to calculate the effect size (95% CI) of Pearson's correlations. The MOPS and individual object presence analyses were performed using R (R Core Team, 2017) and the custom R scripts, also available at https://osf.io/hvgk6/. The *t.test* function was used to perform the independent samples *t*-test with unequal variance and the *p.adjust* function was used to perform the Holm–Bonferroni correction

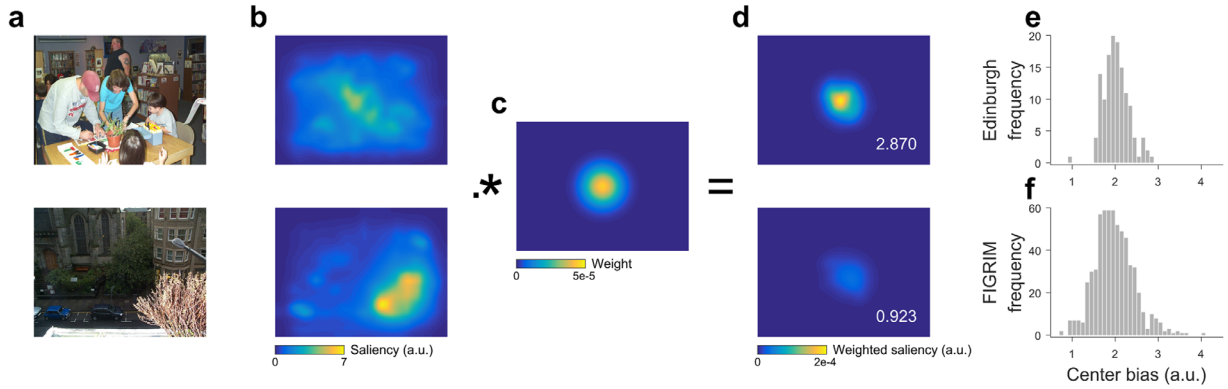

Figure 2. Calculation of center bias. (a) Two exemplar Edinburgh scenes with high and low center bias. (b) GBVS maps. (c) Center bias kernel. (d) The results of element-wise multiplication of parts b and c. (e–f) The distribution of center bias of the Edinburgh and FIGRIM datasets, respectively.

procedure. The *lm* function and *predict* function were used to perform linear regression analyses and predict MOPS in the leave-one-out multivariate regression analysis. The *scale* function was used to *z*-score the data. The *mediate* function with 5000 bootstrap resampling in Psych Package (Revelle, 2020) was used to perform mediation analyses.

Results

Relationships among fixation map consistency, fixation count, and scene memorability in the Edinburgh dataset

We tested whether population-level eye-tracking measures from one group can predict the population-level scene memorability measured from an entirely different group of participants. Specifically, we obtained fixation counts and fixation map consistency measures from group 2 (G2) and used those to predict scene memorability from group 1 (G1; see Supplemental Note S1 for the definition of participant groups G1 and G2).

As a sanity check, we first tested the reliability of the eye-tracking measures across the different groups of participants. We obtained fixation map consistency and fixation counts for each scene from G1 and G2 participants who performed the memorization task (see Methods) and examined the correlation of these measures between G1 and G2. The correlation values were significantly positive for fixation count, Spearman's $\rho(130) = 0.64$; 95% CI, 0.55–0.72; $p < 0.001$, and for fixation map consistency, $\rho(130) = 0.69$; 95% CI, 0.6, 0.75; $p < 0.001$, suggesting that these eye-tracking measures are reliable.

Next, we examined the effects of fixation counts and fixation map consistency on scene memorability by conducting a scene-level linear regression analysis. The dependent variable was recognition accuracy from the G1 participants who performed the memorization task on the scene (scene memorability), and the predictors were fixation map consistency and fixation count, both *z*-scored, from the G2 participants who viewed these scenes on the memorization task. Scene orientation (whether or not a scene was horizontally flipped in the recognition test; see Methods) was also included as a predictor. The correlation plots of the continuous variables are presented in Supplementary Figure S1. Model E_{Both} ($df = 128$), which included both fixation map consistency and fixation count, explained 24.4% of the variance (adjusted R^2). It confirmed significant positive effects of fixation map consistency ($\beta = 0.05$; 95% CI, 0.02–0.08]; $p < 0.001$) and scene orientation ($\beta = -0.17$; 95% CI, –0.23 to −0.11; $p < 0.001$) on scene memorability and a nonsignificant effect of fixation counts ($\beta = 0.02$; 95% CI, –0.01 to 0.05; $p = 0.150$). Figure 3a illustrates these results. Consistent with the linear regression results, the correlation between G2 fixation map consistency and G1 recognition accuracy was significantly positive, $\rho(130) = 0.23$; 95% CI, 0.09–0.36; $p = 0.007$. However, the correlation between G2 fixation counts and G1 recognition accuracy was not significant, $\rho(130) = 0.11$; 95% CI, –0.03 to 0.25; $p = 0.234$.

We then examined whether fixation map consistency and fixation counts differently contribute to scene memorability. The correlation values between fixation map consistency and fixation counts were not significantly different from zero: for G1, $\rho(130) = -0.09$; 95% CI, –0.23 to 0.05; $p = 0.31$; for G2: $\rho(130) = 0.08$; 95% CI, –0.07, 0.22; $p = 0.365$ (Figure 3b). To examine the extent to which these measures can complement in predicting scene memorability, we

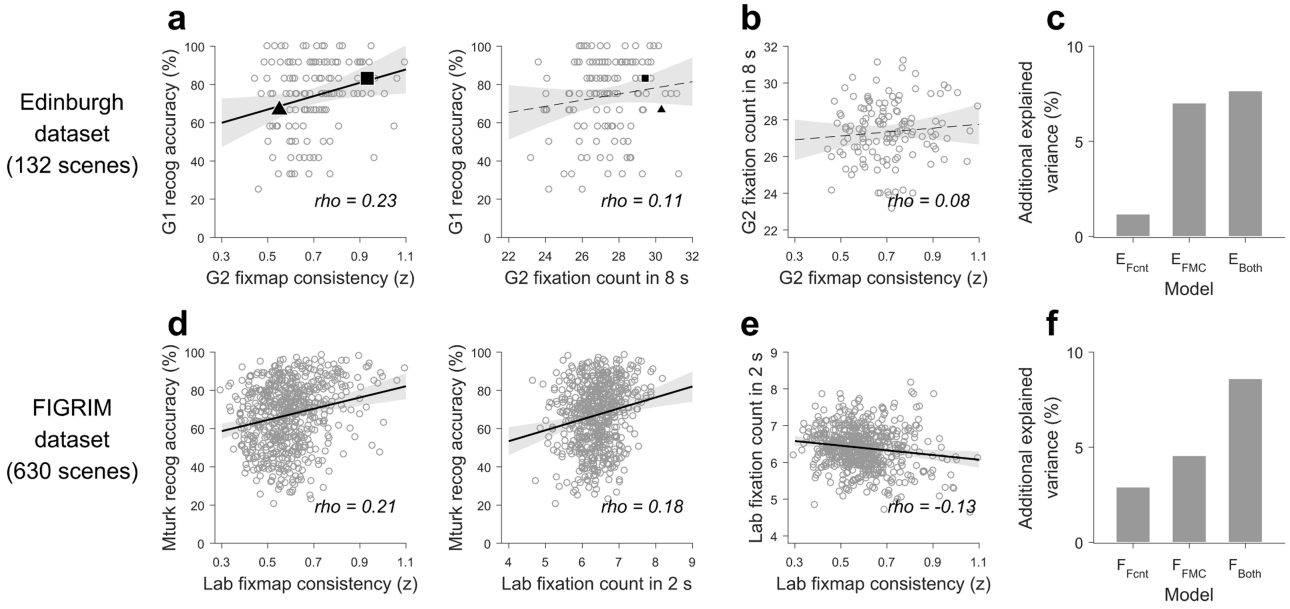

Figure 3. Relationships among fixation map consistency, fixation count, and scene memorability. (a) The Edinburgh results. Scene memorability ("recog accuracy") was obtained from G1. Fixation map ("fixmap") consistency and fixation counts were obtained from G2. Raw values were plotted in the scatterplots. The filled square and triangles indicate the scenes presented in Figure 1. The solid line represents a significant correlation, $\rho(130) = 0.23$; 95% CI, 0.09–0.36; $p = 0.007$. The dashed line represents a nonsignificant correlation, $\rho(130) = 0.11$; 95% CI, –0.03 to 0.25; $p = 0.234$, The gray shades represent the 95% confidence bands. (b) Relationship between fixation map consistency and fixation counts in the Edinburgh dataset. The dashed line represents a nonsignificant correlation, $\rho(130) = 0.08$; 95% CI, –0.07 to 0.22; $p = 0.365$. (c) Explained variance of the linear regression models for predicting scene memorability in the Edinburgh dataset. Models E_{Fcnt} and E_{FMC} used *z*-scored fixation counts and fixation map consistency as the predictor, respectively. Model E_{Both} used both *z*-scored variables as the predictor. (d) The FIGRIM results. Scene memorability was obtained from AMT participants. Fixation map consistency and fixation counts were obtained from the lab participants. The solid lines represent significant correlations. For recognition accuracy and fixation map consistency, $\rho(628) = 0.21$; 95% CI, 0.15–0.27, and for recognition accuracy and fixation count, $\rho(628) = 0.18$; 95% CI, 0.12–0.25 (both $p < 0.001$). (e) Relationship between fixation map consistency and fixation counts in the FIGRIM dataset. The solid line represents a significant correlation, $\rho(628) = -0.13$; 95% CI, –0.2 to –0.07; $p < 0.001$. (f) Explained variance of the linear regression models for predicting scene memorability in FIGRIM dataset. Models F_{Fcnt} and F_{FMC} used *z*-scored fixation counts and fixation map consistency as the predictor, respectively. Model F_{Both} used both *z*-scored variables as the predictor.

conducted scene-level regression analyses using simpler models, where the dependent variable was recognition accuracy from the G1 participants. The base model (Model E_{Base}; $df = 130$) included only scene orientation as a predictor, and we compared it to the models with only fixation counts (Model E_{Fcnt}; $df = 129$), with only fixation map consistency (Model E_{FMC}; $df = 129$), and with both fixation counts and fixation map consistency (Model E_{Both}; $df = 128$). The explained variances were 16.8%, 17.9%, 23.8%, and 24.4% for Models E_{Base}, E_{Fcnt}, E_{FMC}, and E_{Both}, respectively, resulting in an additional 1.2%, 7.0%, and 7.6% of the variance explained by fixation count, fixation map consistency, and both, respectively (Figure 3c). In both Models E_{Fcnt} and E_{Both}, however, including fixation counts did not significantly improve the model fits: $F(2, 129) = 15.31$, $p < 0.001$ and $F(3, 128) = 15.1$, $p < .001$, respectively (Supplementary Table S1).

Relationships among fixation map consistency, fixation count, and scene memorability in the FIGRIM dataset

We repeated the analysis in an independent, larger dataset, the FIGRIM dataset (Bylinskii et al., 2015). We obtained fixation counts and fixation map consistency from the FIGRIM dataset using the same methods as in the Edinburgh dataset, but with one exception; the FIGRIM dataset did not contain fixation duration, so we assigned equal weights for all fixations in generating individual fixation maps. The fixation counts and fixation map consistency of the FIGRIM dataset (viewing duration 2 seconds) were not significantly different from those of the Edinburgh dataset (viewing duration 8 seconds) that were obtained during the first 2 seconds of viewing (Supplementary Note S2; Supplementary Figs. S2a, S2b). We then

Model	Predictor	β	t	p
Model F_{Both}[a] ($df = 606$)	Fixation map consistency	0.043	6.64	<0.001
	Fixation count	0.036	5.62	<0.001
Model F_{Both} + MOPS[b] ($df = 605$)	Fixation map consistency	0.04	6.37	<0.001
	Fixation count	0.031	4.83	<0.001
	MOPS	0.034	4.81	<0.001
Model F_{Both} + face/human, motion[c] ($df = 604$)	Fixation map consistency	0.035	5.74	<0.001
	Fixation count	0.025	3.98	<0.001
	Face/human	0.137	8.46	<0.001
	Motion	0.010	0.55	0.58
Model F_{Both} + MOPS, face/human[d] ($df = 604$)	Fixation map consistency	0.035	5.68	<0.001
	Fixation count	0.024	3.81	<0.001
	MOPS	0.015	2.10	0.036
	Face/human	0.128	7.85	<0.001

Table 1. Contributions of fixation map consistency, fixation count, and MOPS to scene memorability in the FIGRIM dataset. *Notes*: Model comparison results are presented in Supplementary Table S2. Dependent variable: AMT recognition accuracy of 630 scenes. [a]Model F_{Both} included *z*-scored fixation map consistency, *z*-scored fixation count, scene category, and object counts as predictors. [b]The *z*-scored MOPS (see Methods) was added to Model F_{Both}. [c]The presence of face/human (0, absent; 1, present) and the presence of motion in each scene were added to Model F_{Both}. [d]The presence of face/human in each scene and *z*-scored MOPS were added to Model F_{Both}.

conducted a scene-level regression analysis in which the dependent variable was recognition accuracy from the AMT participants (i.e., scene memorability), and the predictors were fixation map consistency and fixation counts of the same scenes, both *z*-scored, from the lab participants. Scene category, which affects scene memorability (Bylinskii, Isola, Bainbridge, Torralba, & Oliva, 2015), was also included as a categorical predictor. We did not include its interaction terms, because scene category did not significantly interact with fixation map consistency or fixation counts in predicting scene memorability (Supplementary Note S3). Also included as a predictor was the number of objects, which was negatively associated with scene memorability: $\rho(628) = -0.09$; 95% CI, –0.16 to –0.03; $p = 0.019$) (Supplementary Figure S3). The correlation plots of the continuous variables are presented in Supplementary Figure S3.

Model F_{Both} ($df = 607$), which included both fixation map consistency and fixation count, explained 18.0% of the variance (adjusted R^2) and showed significant positive effects of both fixation map consistency ($\beta = 0.04$; 95% CI, 0.03–0.06; $p < 0.001$) (Table 1) and fixation counts ($\beta = 0.03$, 95% CI, 0.02–0.05; $p < 0.001$). Figure 3d illustrates these results. Consistent with the linear regression results, the correlation values were significantly positive between recognition accuracy and fixation map consistency, $\rho(628) = 0.21$; 95% CI, 0.15–0.27; $p < 0.001$, and between recognition accuracy and fixation count, $\rho(628) = 0.18$; 95% CI, 0.12–0.25; $p < 0.001$, which were not significantly different from those of the Edinburgh dataset that were obtained during the first 2 second of viewing (Supplementary Note S2; Supplementary Figs. S2d, S2e).

We also examined whether fixation map consistency and fixation counts differently contribute to scene memorability. We found that fixation map consistency and fixation counts were significantly negatively correlated in the FIGRIM dataset (Figure 3e), $\rho(628) = -0.13$, 95% CI, –0.2 to –0.07; $p < 0.001$. Then, we conducted scene-level regression analyses using simpler models, where the dependent variable was AMT recognition accuracy. The base model (Model F_{Base}; $df = 608$) included only scene category as a categorical predictor, and we compared it to the models with only fixation counts and scene category as predictors (Model F_{Fcnt}; $df = 607$) and with only fixation map consistency and scene category as predictors (Model F_{FMC}; $df = 607$). Model F_{Both} contained fixation map consistency, fixation count, and scene category as predictors. The explained variances were 11.09%, 13.99%, 15.64%, and 19.7% for Models F_{Base}, F_{Fcnt}, F_{FMC}, and F_{Both}, respectively, resulting in additional 2.9%, 4.5%, and 8.6% of the variance explained by fixation count, fixation map consistency, and both, respectively (Figure 3f). Further model comparison results are presented in Supplementary Table S2. In both datasets, fixation map consistency better predicted scene memorability than fixation count, and using both eye-tracking measures increased the predictive power.

Examination of scene semantics in the FIGRIM dataset

Scene semantics, such as the presence of these nameable objects, has been shown to affect scene memorability (Isola et al., 2011). Capitalizing on the

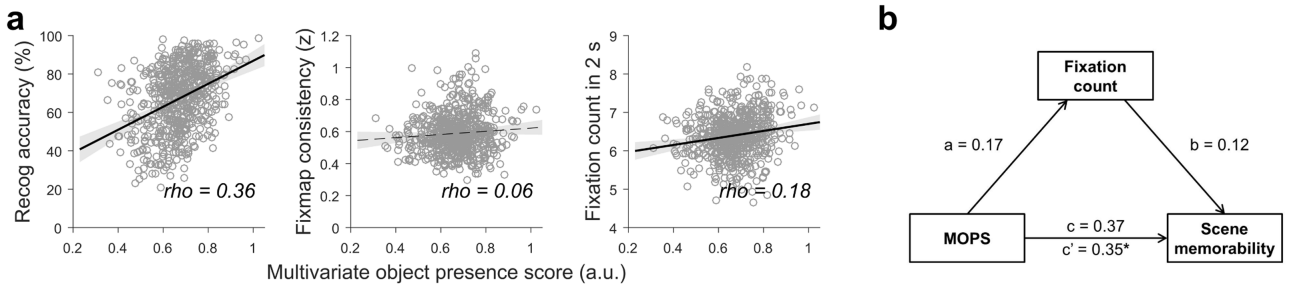

Figure 4. MOPS analysis. (a) The relationships among MOPS, scene memorability (left), fixation map consistency (middle), and fixation counts (right), respectively. The circles represent each image, and the gray shades represent the 95% confidence intervals. Raw values were plotted in the scatterplots. The dashed line represents a non-significant correlation, whereas the solid line represents a significant correlation, with their correlation values written at the right bottom. (b) The parallel mediation model to explain the relationship between the *z*-scored MOPS and scene memorability, with *z*-scored fixation map consistency and *z*-scored fixation counts as mediators.

extensive object annotations in the FIGRIM dataset and using the combined and individual presences of a range of objects in each scene, respectively, as proxies for scene semantics, we asked how object presence contributes to scene memorability. Specifically, we used mediation analysis to examine whether and to what extent eye-tracking measures, such as fixation map consistency and fixation count, could explain the relationship between object presence and recognition accuracy.

First, we created the MOPS as in Isola et al. (2011), in which the presence of 98 selected objects (1 if present, 0 if absent) was weighted-summed to predict scene memorability (see Methods for details about the object selection and leave-one-out MOPS calculation), and we examined its relationships to fixation map consistency and fixation count. As expected, MOPS and AMT recognition accuracy were significantly positively correlated, $\rho(628) = 0.36$; 95% CI, 0.3–0.41; $p < 0.001$ (Figure 4a). We also found that MOPS was significantly positively correlated with and fixation count, $\rho(628) = 0.18$; 95% CI, 0.11–0.24; $p < 0.001$. The relationship between MOPS and fixation map consistency was not significant, $\rho(628) = 0.06$; 95% CI, –0.001 to 0.13; $p = 0.107$. So, we conducted a mediation analysis (see Methods) to investigate the relationships among MOPS, fixation count, and scene memorability (Figure 4b). We found a significant indirect mediation effect ($a \times b$, 0.02; 95% CI, 0.01–0.04) and a significant direct effect, indicating a partial mediation. The results suggest that fixation counts partially mediated the relationship between MOPS and scene memorability.

Second, we inspected all 98 nameable objects and three manually defined semantic features (face/human, motion, watchability) following Xu et al. (2014) (see Methods) individually to identify objects/features that increased scene memorability significantly after correcting for multiple comparisons. Supplementary Table S3 shows the top 20 objects/features in the descending order of mean difference (*t*-score) in

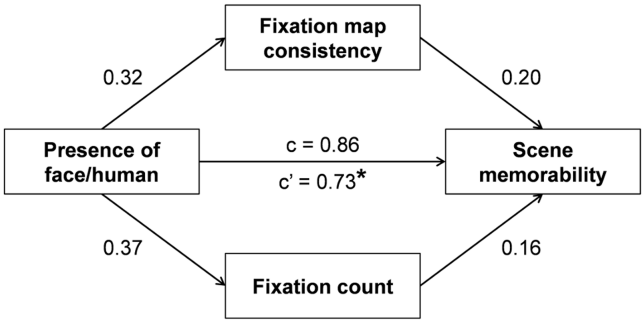

Figure 5. Parallel mediation analysis among face/human, fixation map consistency, fixation count, and scene memorability. This model examined to what extent the presence of face/human was mediated by both *z*-scored fixation map consistency and *z*-scored fixation counts in predicting scene memorability. $^{*}p < 0.05$ (bootstrap test).

scene memorability, obtained with the independent samples *t*-test with unequal variance. We found that the presences of face/human, person, pilot, and motion, all of which are related to face/human, significantly increased scene memorability after correcting for multiple comparisons. We decided to further examine only the presence of face/human because it had the most scenes (158) and mostly overlapped with the presences of person, pilot, and motion (91 out of 101 scenes with person, 10 out of 10 pilot scenes, and 70 out of 119 scenes with motion were coded as face/human). In addition, face/human could explain away the effect of motion on scene memorability (Table 1). We found that the presence of face/human increased both fixation map consistency, $t(229) = -3.21$, $p = 0.002$, and fixation count, $t(246) = -3.86$, $p < 0.001$. After seeing that it was associated with scene memorability, fixation map consistency, and fixation count, we conducted a parallel mediation analysis (Figure 5) to examine whether and to what extent fixation map consistency and fixation counts

could mediate the relationship between face/human and scene memorability. We found a significant total indirect mediation effect ($a \times b$, 0.12; 95% CI, 0.07–0.19) and a significant direct effect, indicating a partial mediation. When individually testing for fixation map consistency path and fixation counts path, respectively, we found significant indirect effects for both fixation map consistency (0.06; 95% CI, 0.02–0.11) and fixation counts (0.06; 95% CI, 0.03–0.11). Together, these results also suggest that both fixation map consistency and fixation counts partly (but not fully) contribute to the relationship between semantic features and scene memorability.

Additive contributions of fixation map consistency, fixation count, and scene semantics to scene memorability in the FIGRIM dataset

To examine whether fixation map consistency, fixation count, and scene semantics additively contribute to scene memorability, we conducted additional scene-level linear regression analyses using MOPS, with the presences of face/human, and the presence of motion as the proxies of scene semantics. Model F_{Both} (Figure 3f) included scene category, fixation map consistency, and fixation counts as predictors and explained 18.0% of the variance in scene memorability (adjusted R^2). By adding z-scored MOPS to Model F_{Both}, the explained variance increased to 22.5%, and the effects of fixation map consistency, fixation count, and MOPS were all significantly positive (Table 1). By adding the presences of face/human and motion to Model F_{Both}, the explained variance increased to 29.1%, and the effects of fixation map consistency, fixation count, and the presences of face/human were significantly positive. However, the effect of motion was not significant, suggesting that it was explained away by face/human. Further model comparison results also demonstrated that adding motion to Model F_{Both} with face/human did not significantly increase model accuracy, $F(1, 604) = 0.30$, $p = 0.584$ (Supplementary Table S2). Finally, by adding both MOPS and the presence of face/human to Model F_{Both}, the explained variance increased to 29.6%, and the effects of fixation map consistency, fixation count, MOPS, and the presence of face/human all remained significantly positive. Because MOPS included the person category in its scoring, adding the presence of face/human in the regression model decreased the effect of MOPS. However, as MOPS also represented other object categories, its effects remained significantly positive after adding face/human (Table 1), and model comparison also confirmed a significant result, $F(1, 604) = 4.42$, $p = 0.036$ (Supplementary Table S2).

To further explicate the relationship between all predictors, we examined the four-way interaction effect of fixation map × fixation count × MOPS × face/human (Supplementary Note S4). We found no interaction effect among all of the variables, indicating their additive contributions in predicting memorability. Together, these results suggest that fixation map consistency, fixation count, and proxies of scene semantics all contribute differently and additively to scene memorability.

Examination of attention deployment across time in the Edinburgh dataset

Fixation map consistency was significantly associated with scene memorability in both the Edinburgh and FIGRIM datasets despite their differences, particularly the viewing duration: 2 seconds in the FIGRIM dataset versus 8 seconds in the Edinburgh dataset. The longer viewing duration in the Edinburgh dataset allowed us to examine the effects of viewing time on attention deployment, such as the temporal consistency of fixation maps across participants and within the same participant, and how these may be related to scene memorability. Specifically, we cut the 8-second fixation data into four 2-second intervals (0–2, 2–4, 4–6, and 6–8 seconds) (Figure 6a) and examined fixation map consistency across G1 and G2 participants in the four intervals (black numbers under the individual fixation maps in Figure 6a, such as G2 FMC: 0.847) and fixation map similarity (i.e., the same Fisher z-transformed Pearson's correlation coefficient used for calculating fixation map consistency; see Methods) between the 0- to 2-second fixation map and the 2- to 4-second, 4- to 6-second, and 6- to 8-second fixation maps within the same participant (blue numbers inside the individual fixation maps in Figure 6a, such as FMSim: 0.651).

We found that fixation map consistency during the interval of 0 to 2 seconds (G1, 0.65 ± 0.15 [$M \pm SD$]; G2, 0.66 ± 0.15) (Figure 6b) was significantly higher than those during the intervals of (1) 2 to 4 seconds (G1, 0.29 ± 0.11; G2, 0.31 ± 0.11); (2) 4 to 6 seconds (G1, 0.23 ± 0.08; G2, 0.24 ± 0.09); and (3) 6 to 8 seconds (G1, 0.20 ± 0.09; G2, 0.25 ± 0.11). This finding suggests that people attend to similar scene regions during the first 2 seconds but attend to different regions afterward. Figure 6a shows this tendency even within the same person. For example, the fixation maps during the intervals of 2 to 4, 4 to 6, and 6 to 8 seconds look very different from that during the 0- to 2-second interval. Consistent with this observation, the similarity values between the 0- to 2-second fixation map and the other fixation maps (i.e., FMSim in Figure 6a) were low (2–4 seconds: G1, 0.18 ± 0.10, G2, 0.17 ± 0.11;

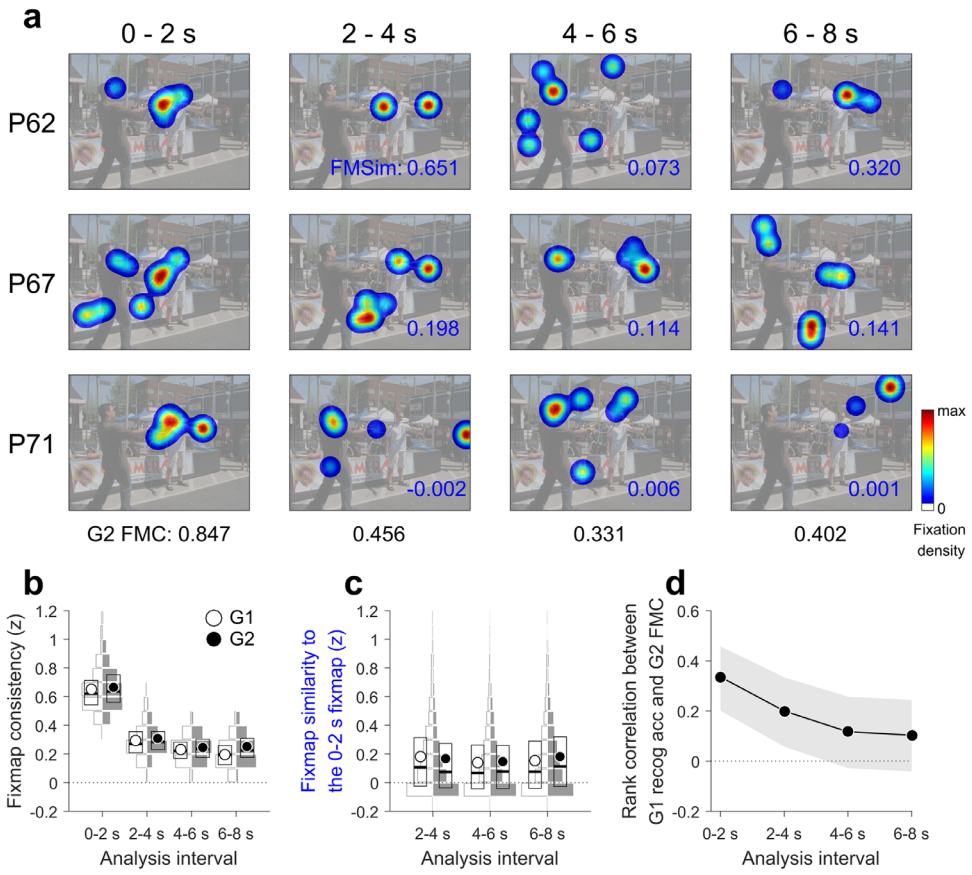

Figure 6. Examination of attention deployment across time in the Edinburgh dataset. (a) Three exemplar G2 participants' fixation maps at each 2-second intervals. The black numbers at the bottom are fixation map consistency (FMC). The blue numbers on the left bottom of each fixation map are fixation map similarity (FMSim) between the current fixation map and the fixation map from 0 to 2 seconds. Note that FMC is calculated across participants, whereas FMSim is calculated within the same participant. (b) FMC across time. The unit of analysis is scene. The histograms of G1 and G2 FMC are drawn, and the rectangles are overlaid to describe their median, first, and third quartiles. The upper and lower horizontal edges of the rectangles specify the first and third quartile, the middle thick lines specify the median, and the open and filled circles represent the mean. (c) FMSim across time. The unit of analysis is trial (a participant seeing a scene), thus the FMSim ranges are larger. However, the majority of FMSim values are around 0, and the mean is around 0.2. (d) Rank correlations between G1 recognition accuracy and G2 FMC across time. The gray shades represent the 95% confidence intervals.

4–6 seconds: G1, 0.14 ± 0.09, G2, 0.15 ± 0.09; 6–8 seconds: G1, 0.15 ± 0.09, G2, 0.18 ± 0.11) (Figure 6c), which contributed to the low levels of fixation map consistency during the intervals of 2 to 4 seconds, 4 to 6 seconds, and 6 to 8 seconds. Moreover, the correlation between fixation map consistency and scene memorability was the highest during the interval of 0 to 2 seconds, $\rho(130) = 0.34$; 95% CI 0.2–0.46; $p < 0.001$ (Figure 6d), and was not a significant predictor of scene memorability after 4 seconds. Together, these results suggest that fixation maps are the most consistent in the first 2 seconds and that the scene features (which we do not fully know yet) that lead to more consistent fixation maps early in viewing may also enhance scene encoding.

Examination of center bias in both datasets

Previous research showed that photographs tend to have objects of interest at their center, resulting in both higher level of low-level visual saliency and higher probability of fixations at the center than in the periphery (Bindemann, 2010; Tatler, 2007; Tseng et al., 2009). It is possible that center bias could affect fixation map consistency and its relationship to scene memorability. For this reason, we calculated the center bias of each scene (Figure 2) and examined the relationships among center bias, fixation map consistency, and recognition accuracy in both datasets.

Comparing center bias and fixation map consistency (Figure 7a), we found that these were significantly

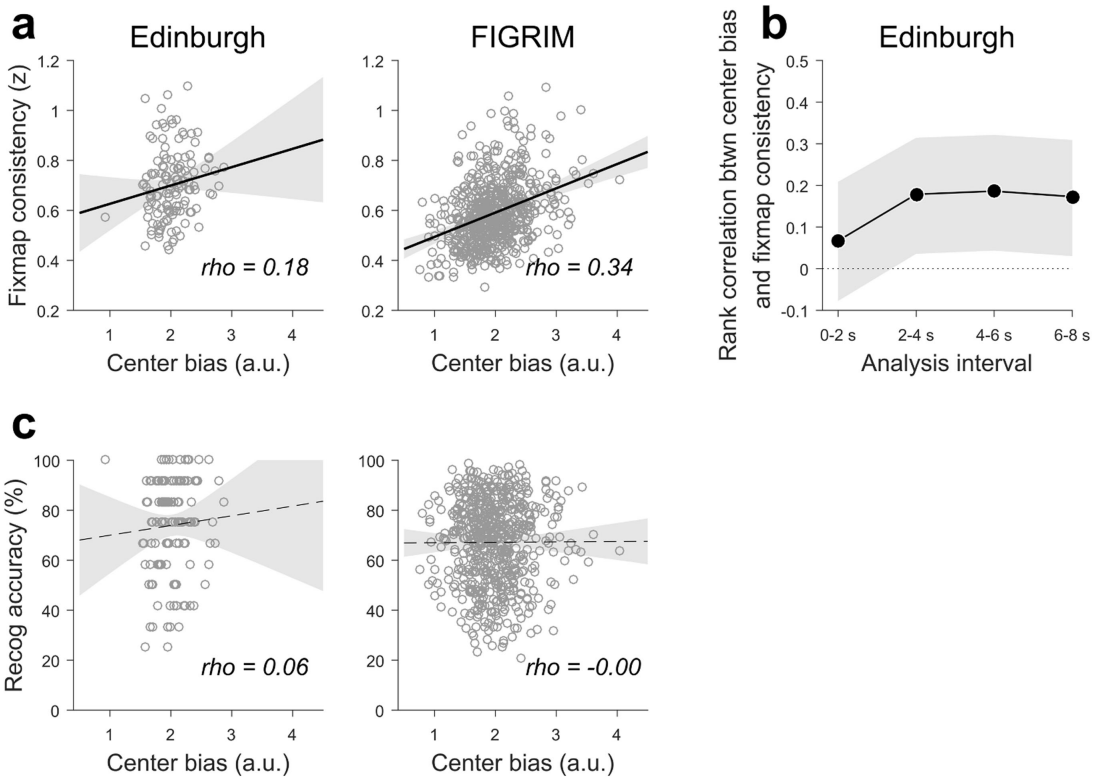

Figure 7. Examination of center bias in both datasets. (a) The relationships between center bias and fixation map consistency in the Edinburgh (left) and FIGRIM (right) datasets. Consistent with the main results (Figure 3), only G2 fixation map consistency is presented. The circles represent each image, and the gray shades represent the 95% confidence intervals. The dashed line represents a non-significant correlation, whereas the solid line represents a significant correlation, with their correlation values written at the right bottom. (b) Correlations between center bias and G2 fixation map consistency across time in the Edinburgh dataset. (c) The relationships between center bias and recognition accuracy in the Edinburgh (left) and FIGRIM (right) datasets.

positively correlated in the FIGRIM dataset, $\rho(628) = 0.34$; 95% CI, 0.28–0.4; $p < 0.001$, and in the Edinburgh dataset, $\rho(130) = 0.18$; 95% CI, 0.04–0.32; $p = 0.036$. In other words, the higher center bias of a scene, the more consistent fixation maps become across participants. The high level of correlation between center bias and fixation map consistency in the FIGRIM raised a possibility that center bias may affect fixation map consistency in the first 2 seconds of viewing (but potentially not later). So, we examined the correlation values between center bias and fixation map consistency across time in the Edinburgh dataset, but we were unable to find a clear relationship between time of viewing and center bias (Figure 7b). Comparing center bias and scene memorability (Figure 7c), we failed to find a significant relationship between center bias and scene memorability in either the Edinburgh dataset, $\rho(130) = 0.06$; 95% CI, –0.08 to 0.2; $p = 0.489$, or the FIGRIM dataset, $\rho(628) = 0.00$; 95% CI, –0.07 to 0.06; $p = 0.970$. Together, these null results suggest that center bias is unlikely to be driving the relationship between fixation map consistency and scene memorability.

Discussion

In this study, we used two different datasets, the Edinburgh and FIGRIM datasets, and confirmed in both datasets that fixation map consistency measured from one group of people was significantly and positively associated with scene memorability measured from a different group of people (Figure 3). Consistent with previous research, we also confirmed the positive effects of fixation counts (Choe et al., 2017; Loftus, 1972; Tatler & Tatler, 2013) and the proxies of scene semantics (Isola et al., 2011) on scene memorability. We found that the relationships between the proxies of scene semantics and scene memorability were partially (but not fully) mediated by fixation map consistency and fixation count, separately as well as together (Figures 4b and 5). Importantly, we found that fixation map consistency, fixation count, and scene semantics additively contributed to scene memorability (Table 1), suggesting that eye tracking can complement computer vision-based algorithms and improve scene memorability prediction.

Although the Edinburgh and FIGRIM datasets were different in many aspects, such as the scenes used, experimental details, and viewing duration, we found that fixation map consistency, fixation count, and their relationships to scene memorability were similar between these datasets (Supplementary Figure S1). In particular, we confirmed a positive association between fixation map consistency and scene memorability in both datasets (Figure 3), consistent with previous research with static scenes (Khosla et al., 2015; Mancas & Le Meur, 2013) and videos (Burleson-Lesser, Morone, DeGuzman, Parra, & Makse, 2017; Christoforou, Christou-Champi, Constantinidou, & Theodorou, 2015). Moreover, the observed correlation values between fixation map consistency and scene memorability from the Edinburgh dataset (0.23; 95% CI, 0.09–0.36) and FIGRIM dataset (0.21; 95% CI, 0.15–0.27) were very similar to the value (0.24) reported by Khosla et al. (2015) from the Fixation Flickr dataset (Judd Ehinger, Durand, & Torralba, 2009). Together, these findings suggest a robust positive association between fixation map consistency and scene memorability.

How can fixation map consistency be associated with scene memorability? In this study, we found three factors, all of which have been reported in previous research, that affected fixation map consistency: center bias, the presence of face/human (which was used as a proxy for scene semantics), and viewing time. How were these factors also related to scene memorability? Regarding center bias (Figure 7), photographs tend to have objects of interest at their center, resulting in both higher level of low-level visual saliency and higher probability of fixations at the center than in the periphery (Bindemann, 2010; Hayes & Henderson, 2020; Tatler, 2007; Tseng et al., 2009). Consistently, we found that center bias was positively correlated with fixation map consistency. However, we also found that center bias was not significantly correlated with scene memorability in the both datasets, suggesting that center bias does not drive the relationship between fixation map consistency and scene memorability.

Regarding the presence of face/human (Figure 5), previous research showed that people prioritize their attention to faces, bodies, and other people (i.e., social features) in naturalistic scenes where they could obtain important social information (Bindemann, Scheepers, Ferguson, 2010; Cerf, Frady, & Koch, 2009; End & Gamer, 2017; Flechsenhar & Gamer, 2017; Scrivner, Choe, Henry, Lyu, Maestripieri, & Berman, 2019), suggesting that scenes with informative and salient scene features, such as faces and people, may produce higher fixation map consistency than scenes without those. For example, Wilming and colleagues (2011) showed that fixation map consistency was higher in urban scenes than in nature scenes and explained that urban scenes have more people and concrete man-made objects, which are more likely to attract fixations. Also, Isola and colleagues (2011) showed that nameable objects, including faces and people, affect scene memorability. Consistently, we found that the presence of face/human simultaneously increased fixation map consistency, fixation count, and scene memorability. Moreover, we found that fixation map consistency and fixation count, separately as well as together, partially mediated its relationship to scene memorability (Figure 5), suggesting that these salient features may enhance scene encoding through multiple mechanisms, including overt visual attention. However, we do not know to what extent nameable objects/features, other than face/human, can engage attentional mechanisms and increase scene memorability because we only examined the objects/features that increased scene memorability significantly after correcting for multiple comparisons in the FIGRIM dataset.

Regarding viewing time (Figure 6), it is well known that fixation maps are more consistent across participants early in viewing (Buswell, 1935; Tatler et al., 2005). Consistently, we found that fixation map consistency was the highest in the first 2 seconds of viewing and quickly dropped afterward (Figure 6b). Moreover, we found the correlation between fixation map consistency and scene memorability was significant only during the intervals of 0 to 2 seconds and 2 to 4 seconds, suggesting that the scene features (other than center bias) that contribute to producing more consistent fixation maps early in viewing may be also important for scene encoding. Unfortunately, however, which scene features can contribute to the difference in fixation map consistency across scenes, especially in the first 2 seconds, is less well known. Our results suggest that such scene features could include highly meaningful features such as faces and people, which can guide overt attention from the very first fixation (Henderson & Hayes, 2017; Henderon & Hays, 2018). Understanding which scene features contribute to producing more consistent fixation maps early in viewing and how these features contribute to scene encoding will be critical for predicting both fixation patterns and scene memorability.

Fixation count is simple, easy to measure, and associated with scene memorability (Choe et al., 2017; Loftus, 1972; Tatler & Tatler, 2013). Consistently, we found that the correlation between fixation counts and scene memorability was significantly positive (0.18; 95% CI, 0.12–0.25) in the FIGRIM dataset and non-significantly positive (0.11; 95% CI, −0.03–0.25) in the Edinburgh dataset, suggesting that the scenes that trigger elaboration are better remembered. Moreover, our finding that fixation counts partially mediated the relationship between MOPS and scene memorability (Figure 4b) also supports the role of elaboration in scene encoding.

Scene semantics is known to play an important role in guiding attention (Cerf et al., 2009; Henderson, 2003; Henderson & Hayes, 2017; Henderson & Hayes, 2018; Wu, Wick, & Pomplun, 2014; Xu et al., 2014) and in forming scene memory (Isola et al., 2011). Consistently, we found that two proxies of scene semantics in this study, MOPS and the presence of face/human, were positively associated with scene memorability. Through the mediation analyses, we found that fixation counts partially mediated the relationship between MOPS and scene memorability (Figure 4b) and that both fixation map consistency and fixation count, separately as well as together, partially mediated the relationship between the presence of face/human and scene memorability (Figure 5). These results suggest that scene semantics engages attentional mechanisms, which contribute to scene encoding, but attentional mechanisms can only partly explain the relationship between scene semantics and scene memorability. As a result, we found that the effects of fixation map consistency, fixation count, MOPS, and the presence of face/human were all significantly positive when these were used together to predict scene memorability (Table 1), suggesting that these measures additively contribute to scene encoding. Our results show why the computer vision-based scene memorability models (Bylinskii et al., 2015; Khosla et al., 2015), which are mainly based on scene information, are successful but also suggest that 2 seconds of eye tracking (for each scene) can provide valuable additional information for better prediction.

There are at least five limitations of this study. First, the scenes used in this study were two-dimensional (2D) computer images and not real scenes, so the results are primarily pertinent to what happens when people look at images on computer screens. Presumably, encoding of 2D computer images would engage similar underlying cognitive processes as encoding real scenes (i.e., three-dimensional), but this assumption has to be explicitly tested. Second, the combined predictive power of the eye-tracking and scene-based measures was still low, as indicated by the explained variance of the full models (Edinburgh E_{Both}: 24.5%; FIGRIM F_{Both} + MOPS + face/human: 30.0%). Third, this study did not fully identify the scene features that can produce high fixation map consistency or provide mechanistic explanations for how a scene can produce more or less consistent fixation maps. Fourth, this study also did not provide mechanistic explanations for how a scene can produce more or less fixations. More fixations in a trial could be interpreted as elaborate inspection (Winograd, 1981), but what and how intrinsic properties of a scene can mechanistically bias fixation counts and saccade rate across viewers have been less studied. Fifth, the proxies of scene semantics used in the study were limited by the existing object annotations and the authors' manual inspection, which may have missed factors affecting scene memorability and attention deployment. Large eye-tracking datasets with a large number of scenes and rapidly improving computer vision algorithms for scene understanding will help tackle these limitations. Future research should further investigate the bottom-up (i.e., scene-specific) and top-down (e.g., instructions, viewing tasks, past experience) factors that affect gaze control (Ballard & Hayhoe, 2009; Henderson, 2007; Henderson, 2011; Henderson, 2017; Tatler, Hayhoe, Land, & Ballard, 2011). Such effort will lead to a precise understanding of what fixation map consistency and fixation counts can tell us about a scene.

Conclusions

By examining two different eye-tracking datasets, we confirmed that the higher the fixation map consistency of a scene, the higher its memorability is. Fixation map consistency and, more importantly, its correlation to scene memorability were the highest in the first 2 seconds of viewing, suggesting that scene features (other than center bias) that contribute to producing more consistent fixation maps early in viewing may be also important for scene encoding. We also found that the relationships between (the proxies of) scene semantics and scene memorability were partially (but not fully) mediated by attentional mechanisms and that fixation map consistency, fixation count, and scene semantics significantly and additively contributed to scene memorability. Together, these results suggest 2 seconds of eye tracking can complement computer vision-based algorithms in better predicting scene memorability.

Keywords: visual attention, image memorability, eye-tracking, fixation map consistency, fixation counts

Acknowledgments

The authors appreciate the editor and the two anonymous reviewers for their thoughtful comments. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Supported by grants from the TFK Foundation and the John Templeton Foundation (University of Chicago Center for Practical Wisdom and the Virtue, Happiness, and Meaning of Life Scholars Group) to M.G.B.; the National Science Foundation to M.G.B. (BCS-1632445); the Mansueto Institute for Urban Innovation to K.W.C. (postdoctoral fellowship); and

the National Eye Institute of the National Institutes of Health to J.M.H. (R01EY027792).

Commercial relationships: none.
Corresponding authors: Kyoung Whan Choe; Marc G. Berman.
Email: kywch@uchicago.edu; bermanm@uchicago.edu.
Address: Department of Psychology, The University of Chicago, Chicago, IL, USA.

*ML and KWC contributed equally to this article.

References

Ballard, D. H., & Hayhoe, M. M. (2009). Modelling the role of task in the control of gaze. *Visual Cognition, 17*(6–7), 1185–1204.

Bindemann, M. (2010). Scene and screen center bias early eye movements in scene viewing. *Vision Research, 50*(23), 2577–2587.

Bindemann, M., Scheepers, C., Ferguson, H. J., & Burton, A. M. (2010). Face, body, and center of gravity mediate person detection in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance, 36*(6), 1477–1485.

Burleson-Lesser, K., Morone, F., DeGuzman, P., Parra, L. C., & Makse, H. A. (2017). Collective behaviour in video viewing: A thermodynamic analysis of gaze position. *PLoS One, 12*(1), e0168995.

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

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research, 116*(Pt B), 165–178.

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision, 9*(3):6, 1–15, https://doi.org/10.1167/9.3.6.

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision, 9*(12):10, 1–15, https://doi.org/10.1167/9.12.10.

Choe, K. W., Blake, R., & Lee, S.-H. (2016). Pupil size dynamics during fixation impact the accuracy and precision of video-based gaze estimation. *Vision Research, 118*, 48–59.

Choe, K. W., Kardan, O., Kotabe, H. P., Henderson, J. M., & Berman, M. G. (2017). To search or to like: Mapping fixations to differentiate two forms of incidental scene memory. *Journal of Vision, 17*(12):8, 1–22, https://doi.org/10.1167/17.12.8.

Christoforou, C., Christou-Champi, S., Constantinidou, F., & Theodorou, M. (2015). From the eyes and the heart: A novel eye-gaze metric that predicts video preferences of a large audience. *Frontiers in Psychology, 6*, 579.

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision, 10*(10):28, 1–17, https://doi.org/10.1167/10.10.28.

Einhäuser, W., & Nuthmann, A. (2016). Salient in space, salient in time: Fixation probability predicts fixation duration during natural scene viewing. *Journal of Vision, 16*(11):13, 1–17, https://doi.org/10.1167/16.11.13.

End, A., & Gamer, M. (2017). Preferential processing of social features and their interplay with physical saliency in complex naturalistic scenes. *Frontiers in Psychology, 8*, 418.

Flechsenhar, A. F., & Gamer, M. (2017). Top-down influence on gaze patterns in the presence of social features. *PLoS One, 12*(8), e0183799.

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems* (pp. 545–552). San Diego, CA: Neural Information Processing Systems Foundation.

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*, 985–994.

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*(4), 219–222.

Henderson, J. M. (2011). Eye movements and scene perception. In S. P. Liversedge, I. D. Gilchrist, & S. Everling (Eds.), *Oxford library of psychology. The Oxford handbook of eye movements* (pp. 593–606). Oxford, UK: Oxford University Press.

Henderson, J. M. (2017). Gaze control as prediction. *Trends in Cognitive Sciences, 21*(1), 15–23.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour, 1*(10), 743–747.

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.

Hollingworth, A. (2012). Task specificity and the influence of memory on visual search: Comment on Võ and Wolfe (2012). *Journal of Experimental Psychology: Human Perception and Performance, 38*(6), 1596–1603.

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 145–152). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *IEEE 12th International Conference on Computer Vision* (pp. 2106–2113). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Kardan, O., Berman, M. G., Yourganov, G., Schmidt, J., & Henderson, J. M. (2015). Classifying mental states from eye movements during scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 41*(6), 1502–1514.

Kardan, O., Henderson, J. M., Yourganov, G., & Berman, M. G. (2016). Observers' cognitive states modulate how visual inputs relate to gaze control. *Journal of Experimental Psychology: Human Perception and Performance, 42*(9), 1429–1442.

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. In *2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 2390–2398). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Le Meur, O., & Baccino, T. (2013) Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behav Res, 45*, 251–266, https://doi.org/10.3758/s13428-012-0226-9.

Loftus, G. R. (1972). Eye fixations and recognition memory for pictures. *Cognitive Psychology, 3*(4), 525–551.

Luke, S. G., Smith, T. J., Schmidt, J., & Henderson, J. M. (2014). Dissociating temporal inhibition of return and saccadic momentum across multiple eye-movement tasks. *Journal of Vision, 14*(14):9, 1–12, https://doi.org/10.1167/14.10.202.

Mancas, M., & Le Meur, O. (2013). Memorability of natural scenes: The role of attention. In *2013 IEEE International Conference on Image Processing* (pp. 196–200). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Nuthmann, A. (2017). Fixation durations in scene viewing: Modeling the effects of local image features, oculomotor parameters, and task. *Psychonomic Bulletin & Review, 24*(2), 370–392.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision, 10*(8):20, 1–19, https://doi.org/10.1167/10.8.20.

Olejarczyk, J. H., Luke, S. G., & Henderson, J. M. (2014). Incidental memory for parts of scenes from eye movements. *Visual Cognition, 22*(7), 975–995.

Pajak, M., & Nuthmann, A. (2013). Object-based saccadic selection during scene perception: evidence from viewing position effects. *Journal of Vision, 13*(5):2, 1–21, https://doi.org/10.1167/13.5.2.

Pomplun, M., Ritter, H., & Velichkovsky, B. (1996). Disambiguating complex visual information: Towards communication of personal views of a scene. *Perception, 25*(8), 931–948.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramey, M. M., Henderson, J. M., & Yonelinas, A. P. (2020). The spatial distribution of attention predicts familiarity strength during encoding and retrieval. *Journal of Experimental Psychology: General,* https://doi.org/10.1037/xge0000758.

Revelle, W. R. (2020). psych: Procedures for psychological, psychometric, and personality research. Retrieved from https://cran.r-project.org/web/packages/psych/index.html.

Scrivner, C., Choe, K. W., Henry, J., Lyu, M., Maestripieri, D., & Berman, M. G. (2019). Violence reduces attention to faces and draws attention to points of contact. *Scientific Reports, 9*(1), 17779.

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):4, 1–17, https://doi.org/10.1167/7.14.4.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research, 45*(5), 643–659.

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, 1–23, https://doi.org/10.1167/11.5.5.

Tatler, B. W., & Tatler, S. L. (2013). The influence of instructions on object memory in a real-world setting. *Journal of Vision, 13*(2):5, 1–13, https://doi.org/10.1167/13.2.5.

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*(4), 766–786.

Tseng, P.-H., Carmi, R., Cameron, I. G. M., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic

natural scenes. *Journal of Vision, 9*(7):4, 1–16, https://doi.org/10.1167/9.7.4.

Wilming, N., Betz, T., Kietzmann, T. C., & König, P. (2011). Measures and limits of models of fixation selection. *PLoS One, 6*(9), e24038.

Winograd, E. (1981). Elaboration and distinctiveness in memory for faces. *Journal of Experimental Psychology: Human Learning and Memory, 7*(3), 181–190.

Wolfe, J. M., Horowitz, T. S., & Michod, K. O. (2007). Is visual attention required for robust picture memory? *Vision Research, 47*(7), 955–964.

Wooding, D. S. (2002). Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps. *Behavior Research Methods, Instruments, & Computers, 34*(4), 518–528.

Wu, C.-C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology, 5*, 54.

Wyatt, H. J. (2010). The human pupil and the use of video-based eyetrackers. *Vision Research, 50*(19), 1982–1988.

Xu, J., Jiang, M., Wang, S., Kankanhalli, M. S., & Zhao, Q. (2014). Predicting human gaze beyond pixels. *Journal of Vision, 14*(1):28, 1–20, https://doi.org/10.1167/14.1.28.