NON PEER-REVIEWED PREPRINT

Scenes that produce more consistent fixation maps are more memorable

Muxuan Lyu[1]†, Kyoung Whan Choe[1,2]†*, Omid Kardan[1], Hiroki P. Kotabe[1], John M. Henderson[3], & Marc G. Berman[1,4]*

[1]Department of Psychology, The University of Chicago; [2]Mansueto Institute for Urban Innovation, The University of Chicago; [3]Center for Mind and Brain and Department of Psychology, University of California, Davis; [4]Grossman Institute for Neuroscience, Quantitative Biology and Human Behavior, The University of Chicago

† These authors contributed equally to this work.

* Corresponding authors:

- Kyoung Whan Choe, Department of Psychology, The University of Chicago, 5848 S. University Avenue, Chicago, IL 60637. E-mail: kywch@uchicago.edu
- Marc G. Berman, Department of Psychology, The University of Chicago, 5848 S. University Avenue, Chicago, IL 60637. E-mail: bermanm@uchicago.edu

Commercial relationships: none.

Abstract

Studying factors that contribute to scene memorability is important for understanding human vision and memory. Here we demonstrated in two different eye-tracking datasets that the higher the fixation map consistency (also called inter-observer congruency of fixation maps) of a scene, the higher its memorability is. To provide a mechanistic explanation for how a scene can produce more or less consistent fixation maps across viewers, we created a simple computational model by assuming some high signal regions in a scene that will attract more fixations than other regions (ambient noise). We then varied the amplitude of the signal relative to noise (SNR) to examine the relationship between SNR and fixation map consistency. Our model showed that the higher a scene's SNR, the higher its fixation map consistency, suggesting that fixation map consistency reflects the SNR of a scene, an intrinsic scene property that can affect human vision and memory.

Keywords: Visual Attention, Scene Memorability, Eye-tracking, Fixation Map Consistency, Fixation Count

Introduction

Some visual scenes are more memorable than other scenes (Isola, Xiao, Torralba, & Oliva, 2011). Investigating scene memorability is not only important for understanding human vision and memory but is also useful for people interested in predicting and maximizing it for practical purposes. Previous research has shown that intrinsic features of a scene, such as global descriptors, objects counts or areas, semantic features, interestingness, and aesthetics, can affect scene memorability in a similar manner across different viewers (Isola et al., 2011). 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, Torralba, & Oliva, 2015) and the viewing tasks that were performed while each scene was presented (Choe, Kardan, Kotabe, Henderson, & Berman, 2017). Despite great research interest, factors in a scene that could contribute to its memorability are not fully understood.

Eye-tracking enables the investigation of underlying attentional mechanisms of scene memory by measuring fixation counts and fixation maps (Henderson, 2003; Pomplun, Ritter, & Velichkovsky, 1996; Wooding, 2002), i.e., where viewers look in scenes. First, an increased fixation count during encoding is associated with better recognition for scenes (Choe et al., 2017) and objects (Tatler & Tatler, 2013) on a trial-by-trial basis, suggesting that fixation count signals viewers' elaborate inspection of a scene and that elaborate inspection can enhance scene encoding (Winograd, 1981). Second, more preferred scenes produce more fixations and are better remembered later than less preferred scenes (Loftus, 1972), suggesting that the averaged fixation count across viewers can reveal an intrinsic property of a scene, such as interestingness (e.g., the more interesting a scene is, the more elaborate inspection viewers do). Finally, where viewers attend to during scene encoding affects intentional and incidental scene memory (Choe et al., 2017; Hollingworth, 2012; Olejarczyk, Luke, & Henderson, 2014; Tatler & Tatler, 2013). For example, Choe and colleagues (2017) showed that the fixation map from a scene while participants intentionally memorized the scene was different from the fixation maps when participants were searching for an object. Additionally, the degree of difference in the fixation maps for visual search vs. scene memorization in the same scene could explain how and why visual search impaired incidental scene memory on a trial-by-trial basis.

Fixation map consistency (Dorr, Martinetz, Gegenfurtner, & Barth, 2010; Torralba, Oliva, Castelhano, & Henderson, 2006), i.e., the consistency of fixation maps across viewers (also called inter-observer congruency or inter-subject consistency), is a scene-specific, population measure (i.e., averaged over a population of participants for each scene), which is 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). One very interesting, but less explored question arises: how does a scene produce more or less consistent fixation maps across viewers? It is well established that some semantic visual features such as faces strongly attract attention and fixations (Cerf, Frady, & Koch, 2009; Henderson & Hayes, 2017, 2018; Kardan et al., 2017; Xu, Jiang, Wang, Kankanhalli, & Zhao, 2014). Moreover, the scene regions that consistently attract fixations across viewers have been rated as more informative (Mackworth & Morandi, 1967; McCamy, Otero-Millan, Di Stasi, Macknik, & Martinez-Conde, 2014), more interesting (Einhäuser & Nuthmann, 2016; Masciocchi, Mihalas, Parkhurst, & Niebur, 2009; Onat, Açık, Schumann, & König, 2014), and more meaningful (Henderson & Hayes, 2017, 2018). 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. Thus, it is plausible that scenes with features/regions that can strongly attract fixations would be better remembered and produce more consistent fixation maps across viewers.

Importantly, two recent papers have shown that the fixation map consistency is positively associated with scene memorability (Khosla, Raju, Torralba, & Oliva, 2015; Mancas & Le Meur, 2013). However, those studies did not simultaneously examine the effect of fixation count, which is also associated with scene memory (Choe et al., 2017; Loftus, 1972). Therefore, it is important to investigate whether and how fixation map consistency uniquely contributes to scene memory that is different from fixation count. In addition, to understand the mechanisms how fixation map consistency is related to scene memory, one should understand what information fixation map consistency provides us about a scene. Thus, we created a simple computational model that can explain how a scene can produce consistent fixation maps.

We utilized two previous 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 test whether population-level eye-tracking measures from one group can predict population-level scene memory measured from an entirely different group of participants. One advantage of using population measures is that these measures reflect intrinsic properties of a scene that could be easily applied to a new population. By examining the relationship between scene memory, fixation count, and fixation map consistency, we found that fixation map consistency was reliably and significantly associated with scene memorability across the two different datasets, replicating previous research (Khosla et al., 2015; Mancas & Le Meur, 2013). In addition, we found that fixation map consistency and fixation count were not significantly positively correlated, suggesting that fixation map consistency uniquely contributes to scene memory. From these results we then constructed a computational model for how consistent fixation maps could emerge for a scene, thereby allowing us to speculate quantitatively on what fixation map consistency measures about a scene.

Methods

Overview

This study is a reanalysis of two previously collected eye-tracking datasets, the sample sizes of which were determined for different purposes. First, we performed exploratory analyses on the Edinburgh dataset (Luke et al., 2014; Nuthmann & Henderson, 2010; Pajak & Nuthmann, 2013), which 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 (i.e., intentional memorization, visual search, and aesthetic preference evaluation) from 72 participants and scene memorabiliy scores of these scenes from a subset of the participants (36). Scene memorability score in the Edinburgh dataset was the averaged recognition accuracy in a subsequent surprise memory test, i.e., the number of hit (correctly recognized) trials divided by the number of hit and miss trials (which equaled the number of participants (36) who saw these scenes). Out of 135 scenes, we only analyzed the 132 scenes that were used in both the encoding tasks and memory test. Also, we limited analyses to the intentional memorization task from the 24 participants who performed this task on the 132 scenes. These data produced 24 fixation maps per scene, and recognition accuracy from 12 participants per scene. Please see the supplementary method for details.

We then performed a confirmatory analysis on the FIGRIM dataset (Bylinskii et al., 2015), which is freely available at http://figrim.mit.edu/index_eyetracking.html. This dataset used a continuous scene recognition task (Isola et al., 2011), where a series of scenes were presented to participants for encoding. Some of those scenes were later repeated to test recognition memory, where participants were asked to detect repeated scenes. This dataset included fixation maps and memorability scores from 630 scenes coming from 21 different scene categories (30 images per category). These eye-tracking data were generated from 67 in-lab participants (16 participants per scene) and the scene memorability score came from 74 Amazon Mechanical Turk workers. Scene memorability score in the FIGRIM dataset was the proportion of hit trials when a scene was shown the second time to test its memory, i.e., the number of hit (correctly recognized) trials divided by the number of hit and miss trials (which equaled the number of participants who saw these scenes twice). For more methodological details of the FIGRIM dataset, see Bylinkskii et al. (2015). All of our analysis codes are available at https://osf.io/hvgk6/.

Eye movement analysis

Edinburgh dataset. The raw eye movement data were preprocessed using Eyelink Data Viewer (SR Research) to identify discrete fixations and fixation durations during 8 s of scene viewing. Saccades were defined with a 50°/s velocity threshold using a 9-sample saccade detection model. 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. The *fixation count* is the number of discrete fixations, regardless of their duration, that landed on the scenes.

FIGRIM dataset. Bylinskii et al. "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)" (Bylinskii et al., 2015). Note that the FIGRIM dataset does not contain fixation duration information. The fixation count is the number of discrete fixations on the scenes.

Fixation map analysis. We performed analyses using custom MATLAB (MathWorks, Natick, MA, USA) scripts. An individual fixation map of a participant viewing a scene (Fig. 1) was constructed by convolving a Gaussian kernel over its duration-weighted fixation locations during 8 s of viewing (the Edinburgh dataset) or over equal-weighted fixation locations during 2 s of viewing (the FIGRIM dataset). The full width at half maximum of the Gaussian kernel was set to 2° (i.e., σ=0.85°) to simulate central foveal vision and to take into account the measurement errors of video-based eye trackers (Choe, Blake, & Lee, 2016).

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 (i.e., 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 Fig. 1b, twelve similarity values were obtained by comparing each individual fixation map vs. the average of the other eleven fixation maps; then those twelve values were averaged to produce a fixation map consistency score. For the similarity metric, we opted for the Fisher z-transformed Pearson 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.

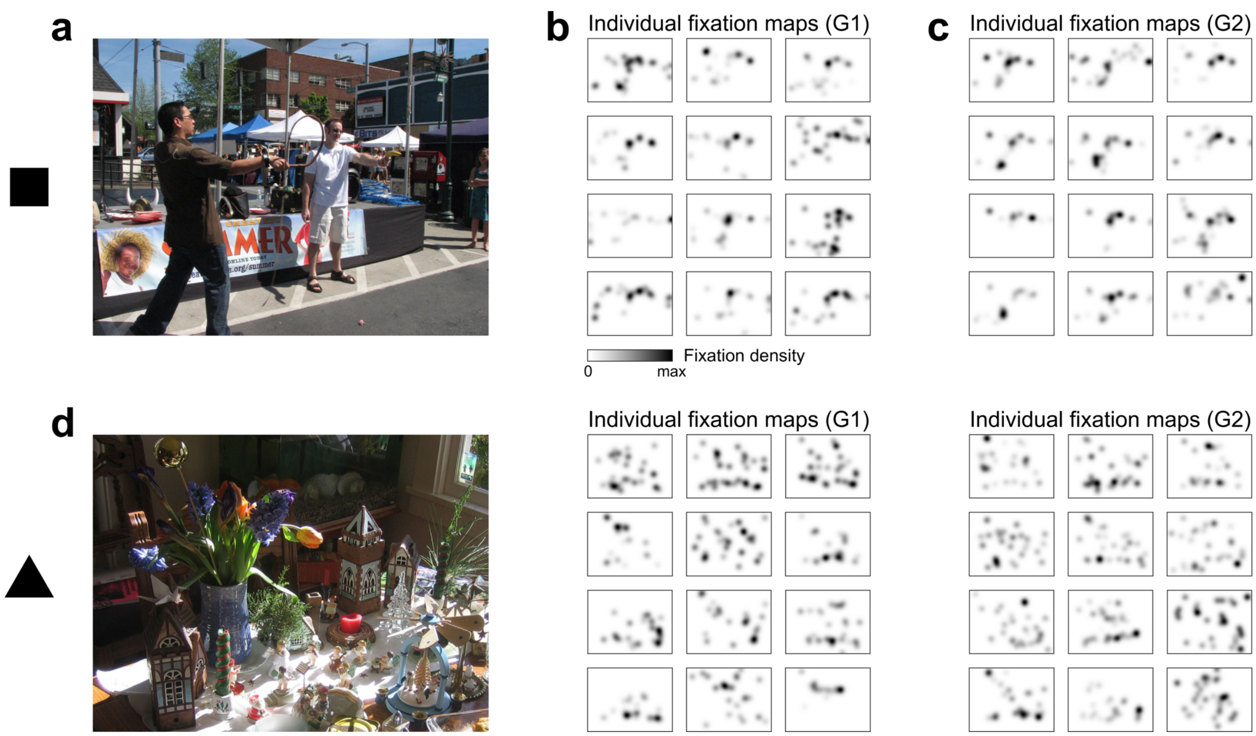

Fig. 1. Individual fixation maps of two example scenes during intentional memorization from the Edinburgh dataset. (**a**) A scene that produced highly consistent fixation maps. The filled square in the left are used in Figs. 2 and 3e to indicate this scene. (**b**) Fixation maps of twelve G1 participants, who were asked to memorize the scene and tested for their scene

memory in the following recognition test (see Supplemental Method). The average recognition accuracy of these participants was used as the scene memorability. (**c**) Fixation maps of twelve G2 participants, who were asked to memorize the scene but *not* tested for their memory. (**d**) Individual fixation maps from G1 (the middle panel) and G2 (the right panel) for another scene (the left panel) that produced less consistent fixation maps. The triangle in the left is in Figs. 2 and 3e to indicate this scene.

Simulation

To examine how a scene can produce consistent fixation maps across viewers, we performed a Monte Carlo simulation of fixation generations. In the simulation, a 1-dimensional screen of 800 pixels (matching the image width of the Edinburgh dataset) was used, and the locations of seven fixations (2 seconds of viewing results in about seven fixations) were randomly generated on the screen for each virtual participant, assuming 2 seconds of scene viewing. The probability density function (PDF) that was used to generate random fixations of virtual participants, which was the same across the participants, consisted of uniformly-distributed noise and a Gaussian signal of σ=31 pixels at the screen center; the amplitude of the signal was set relative to that of uniform noise, thus allowing us to manipulate signal-to-noise ratio (SNR) for investigation. We varied SNR from 0 to 15 in increments of 1 and simulated 2000 sets of 12 virtual participants (matching the number of participants in the Edinburgh dataset) for each SNR (16 SNRs x 2000 sets = 32000 sets). Each simulation set was created by randomly generating seven fixations for each participant using the PDF associated with that SNR. Next, we generated individual fixation maps by smoothing the fixations using the same-sized Gaussian kernel as we did in the Edinburgh dataset (i.e., the full width at half maximum of 2° or 62 pixels). Then, fixation map consistency was calculated for each set of twelve participants using the same leave-one-out procedure.

To examine the number of viewers necessary to obtain reliable fixation map consistency, we repeated the above Monte Carlo simulation with different numbers (6, 24, and 48) of virtual participants in a set and generated an additional 96000 sets (3 numbers of participants x 2000 simulation sets x 16 SNRs).

Statistical software

The fixation map analyses and Monte Carlo simulation were performed using custom MATLAB scripts, available at https://osf.io/hvgk6/. All statistical tests were performed using

MATLAB R2015b. 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 random fixations of virtual participants were generated from the fixation PDFs using the *rand_generator* function, which was obtained from https://www.mathworks.com/matlabcentral/fileexchange/40598.

Results

Exploratory analyses in the Edinburgh dataset

We tested whether population-level eye-tracking measures from one group can predict the population-level scene memory measured from an entirely different group of participants. Specifically, we obtained fixation count and fixation map consistency measures (see Method) for each scene from G2 and used those to predict scene memorability from G1.

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 scores and averaged fixation counts for each scene from G1 and G2 participants who performed the memorization task and examined the correlation of these measures across the different groups. The correlation values were significantly positive for fixation count, Pearson's $r(130) = 0.69$, 95% CI [0.58, 0.77], $p < .001$, and fixation map consistency, $r(130) = 0.67$, 95% CI [0.56, 0.75], $p < .001$, suggesting that these eye-tracking measures are reliable.

To examine the effects of fixation count and fixation map consistency on scene memory, we conducted a scene-level linear regression analysis. The dependent variable was the average recognition accuracy from the G1 participants who performed the memorization task on the scene, and the predictor variables were fixation map consistency scores and averaged fixation counts, both z-scored, from the G2 participants who viewed these scenes on the memorization task. Scene orientation (i.e., whether a scene was horizontally flipped in the recognition test; see Methods) was also included as a predictor. The Edinburgh full model (Model Ef; $df = 128$) explained 24.4% of the variance (adjusted R^2). The Model Ef (i.e., Edinburgh full) 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, -0.11], $p < 0.001$) on scene memory and a nonsignificant effect of fixation count ($\beta = 0.02$, 95% CI [−0.01, 0.05], $p = 0.15$). Fig. 2a illustrates these results. Consistent with the linear regression results, the correlation between G2 fixation map consistency and G1 recognition accuracy was significantly positive, $r(130) = 0.25$, 95% CI [0.08, 0.40], $p = 0.004$, but the correlation between G2 fixation count and G1 recognition accuracy was not significant, $r(130) = 0.14$, 95% CI [−0.04, 0.30], $p = 0.12$, but was in the positive direction of predicting higher recognition accuracy.

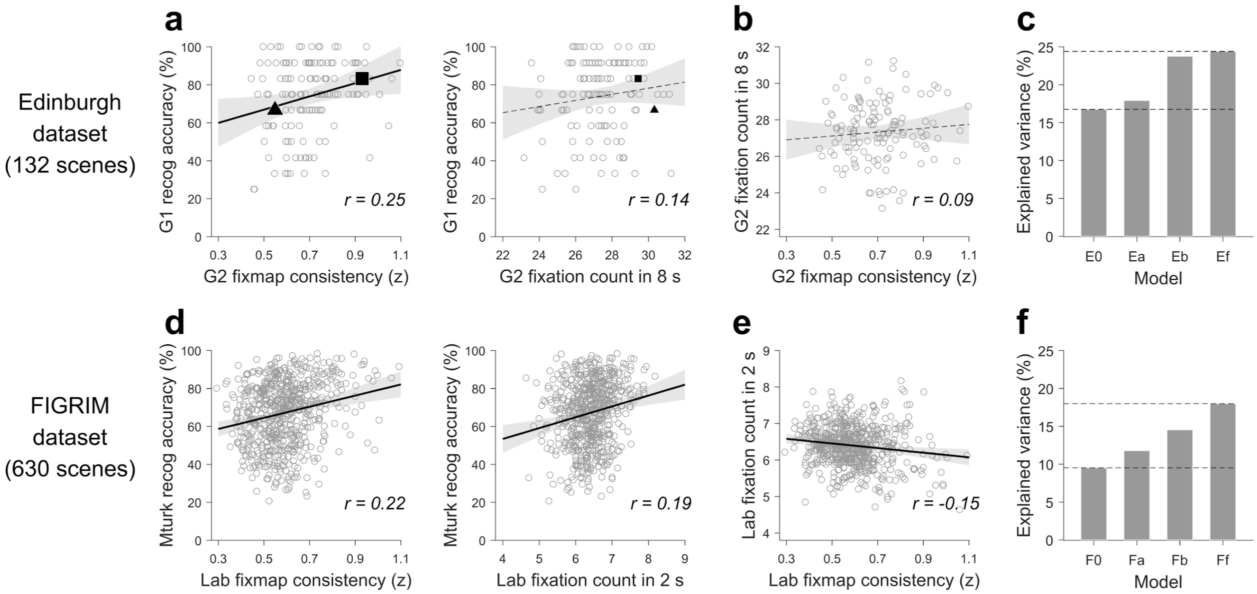

Fig. 2. Relationships between fixation map consistency, fixation count, and scene memorability. (a) The Edinburgh results. Scene memorability (i.e., "recog accuracy") was obtained from G1. Fixation map (i.e., "fixmap") consistency and fixation count were obtained from G2. The filled square and triangles indicate the scenes presented in Fig. 1. The solid line represents a significant linear regression, $r(130) = 0.25$, 95% CI [0.08, 0.40], $p = 0.004$, the dashed line represents a nonsignificant regression, $r(130) = 0.14$, 95% CI [−0.04, 0.30], $p = 0.12$, and the gray shades represent the 95% confidence bands. (b) Relationship between fixation map consistency and fixation count in the Edinburgh dataset. The dashed line represents a nonsignificant regression, $r(130) = 0.09$, 95% CI [−0.08, 0.26], $p = 0.32$. (c) Explained variance of the linear regression models for predicting scene memorability. Two horizontal dashed lines represent the values of Edinburgh base (E0) and full (Ef) models. The Ea and Eb models used only either fixation count or fixation map consistency, respectively. (d) The FIGRIM results. Scene memorability was obtained from Amazon Mechanical Turk participants. Fixation map consistency and fixation count were obtained from the lab participants. The solid lines represent significant regressions: $r(628) = 0.22$, 95% CI [0.14, 0.29]; $r(628) = 0.19$, 95% CI [0.11, 0.26]; both ps < 0.001. (e) Relationship between fixation map consistency and fixation count in the FIGRIM dataset. The solid line represents a significant regression, $r(628) = -0.15$, 95% CI [−0.22, -0.07], $p < 0.001$. (f) Explained variance of the linear regression models for predicting scene memorability. Two horizontal dashed lines represent the values of FIGRIM base (F0) and full (Ff) models. The Fa and Fb models used only either fixation count or fixation map consistency, respectively.

We also examined whether and how fixation map consistency uniquely contributes to scene memory that is different from fixation count. We found that the correlation values between fixation map consistency and fixation count were not significantly different from zero (G1: $r(130) = -0.09$, 95% CI [−0.26, 0.08], $p = 0.29$; G2: $r(130) = 0.09$, 95% CI [−0.08, 0.26], $p = 0.32$; Fig. 2b), suggesting these factors encode different information and thus increase predictive power when used together. To examine the extent to which these measures can complement in predicting scene memory, we conducted scene-level regression analyses using simpler models, where the dependent variable was the average recognition accuracy from the G1 participants. The base model (Model E0; $df = 130$) only included scene orientation as a predictor. Also, we examined the models with only fixation count (Model Ea; $df = 129$), with only fixation map consistency (Model Eb; $df = 129$), and with both fixation count and fixation map consistency (Model Ef; df = 128). The explained variances were 16.8%, 17.9%, 23.8%, and 24.4% for Models E0, Ea, Eb, and Ef respectively (Fig. 2c). These variables resulted in an additional 1.1%, 7.0%, and 7.6% of the variance explained by fixation count, fixation map consistency, and both, respectively.

Confirmatory analyses in the FIGRIM dataset

To see if the results from the Edinburgh dataset could be generalized, we repeated the analysis in an open-access and larger dataset; the FIGRIM dataset (Bylinskii et al., 2015). This dataset contains 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. 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. The memorability score for each scene was defined as the number of hit trials (i.e., trials where participants correctly pressed a button to a repeat scene) divided by the sum of both hit trials and miss trials (i.e., trials where participants did not press a button to a repeat scene) across participants.

We calculated the averaged fixation count 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, thus we assigned equal weights for all fixations in generating individual fixation maps. Then, we conducted a scene-level linear regression analysis, in which the dependent variable was scene memorability scores from the AMT participants, and the predictors were fixation map consistency and the averaged fixation count of the same scenes, both *z*-scored, from the lab participants. Scene category was also included a categorical predictor. The FIGRIM full model (Model Ff; $df = 607$) explained 18.0% of the variance (adjusted R^2). The Model Ff (i.e., FIGRIM full) showed significant positive effects of both fixation map consistency ($\beta = 0.04$, 95% CI [0.03, 0.06], $p < 0.001$) and fixation count ($\beta = 0.03$, 95% CI [0.02, 0.05], $p < 0.001$). Fig. 2d illustrates these results. Consistent with the linear regression results, the correlation values were both significantly positive between the in-lab fixation map consistency scores and AMT scene memorability, $r(628) = 0.22$, 95% CI [0.14, 0.29], $p < 0.001$, and between the in-lab fixation count and AMT scene memorability, $r(628) = 0.19$, 95% CI [0.11, 0.26], $p < 0.001$.

To examine whether and how fixation map consistency uniquely contributes to scene memory that is different from fixation count, we performed similar correlation and regression analyses as we did in the Edinburgh dataset. In the FIGRIM dataset, we found that the correlation values between fixation map consistency and fixation count were significantly negatively correlated (Fig. 2e), $r(628) = -0.15$, 95% CI [−0.22, -0.07], $p < 0.001$. Then, we conducted scene-level regression analyses using using simpler models, where the dependent variable was the scene memorability scores from the AMT participants. The base model (Model F0; $df = 609$) only included scene category as a categorical predictor. Also, we examined the models with only fixation count (Model Fa; $df = 608$) and with only fixation map consistency (Model Fb; $df = 608$), which included fixation count and fixation map consistency, respectively, as a predictor as well as scene category. The full model, Ff, contained both fixation map consistency and fixation count as well as scene category as predictors. The explained variances were 9.5%, 11.8%,14.5%, and 18% for Models F0, Fa, Fb, and Ff respectively (Fig. 2f), resulting in additional 2.3%, 5.0%, and 8.5% variance of the explained by fixation count, fixation map consistency, and both, respectively. In both datasets, fixation map consistency better predicted scene memorability than fixation count, and using both eye-tracking measures yielded the highest predictive power. In FIGRIM, fixation count was a stronger predictor of memorability than in the Edinburgh dataset, but in both datasets fixation map consistency was a stronger predictor.

Comparisons between the Edinburgh and FIGRIM datasets

Fixation map consistency was found to be significantly associated with scene memory in both the Edinburgh and FIGRIM datasets, but fixation count was found to be significant only in the FIGRIM dataset. Why do these results differ? There are several differences between those two datasets, such as the scene stimuli, experimental paradigms, and participants. However, one notable difference was the viewing duration: 8 seconds in the Edinburgh dataset vs. 2 seconds in the FIGRIM dataset. Therefore, we examined the effects of viewing duration on fixation map consistency, fixation count, and their relationship to scene memory. Specifically, we analyzed the Edinburgh data by varying the analysis from analyzing 1 second of eye-tracking data to the full 8 seconds of viewing in 1 second increments (the filled circles in Fig. 3). We then compared the Edinburgh results at 2 seconds (i.e., the fixations within the first 2 seconds were analyzed) with the FIGRIM results (the filled stars in Fig. 3).

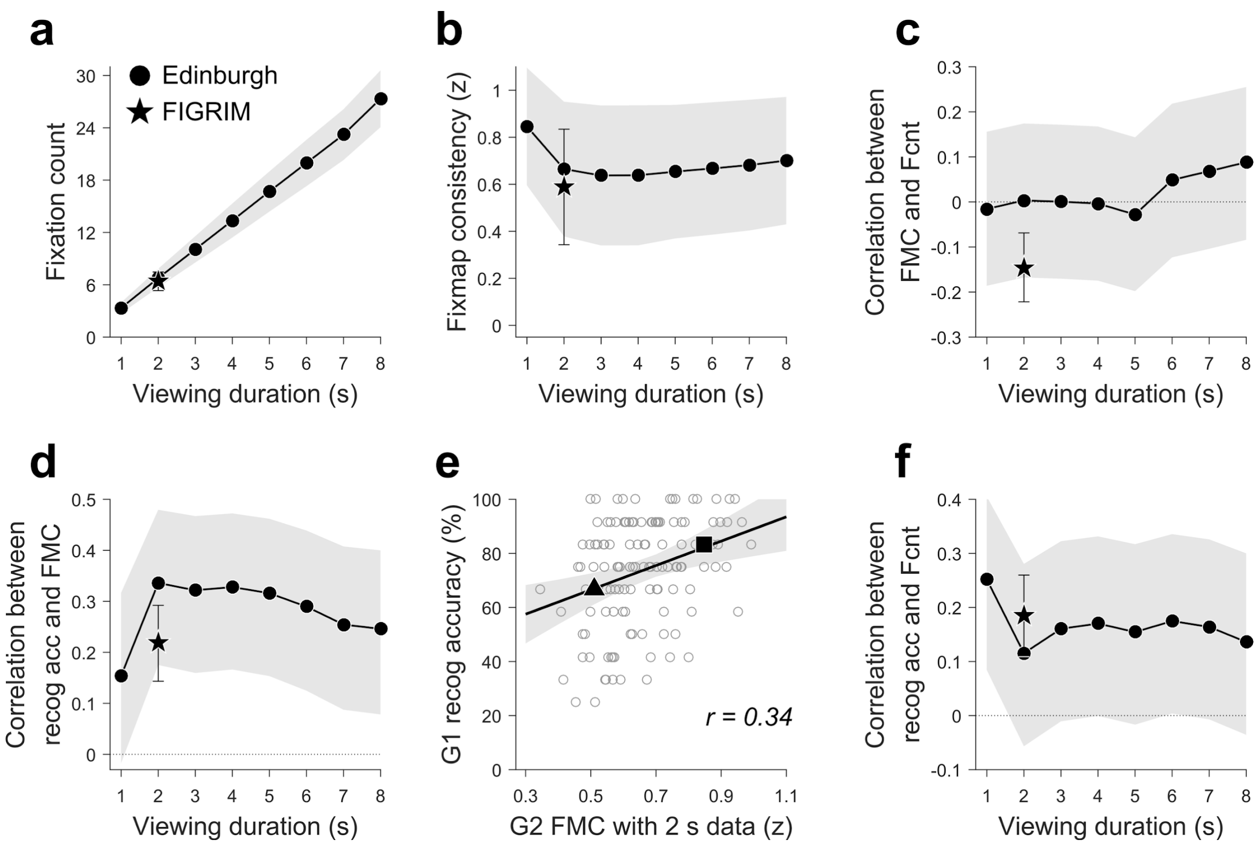

Fig. 3. Relationships between fixation map consistency (FMC), fixation count (Fcnt), and scene memory (recog acc) with respect to viewing duration. The viewing duration of 4

seconds indicates that the first 4 seconds (from the trial onset) of the fixation data were used to calculate fixation map consistency and fixation count. (**a**) Fixation count over viewing duration. The filled circles represent the Edinburgh results, the filled stars represent the FIGRIM results, the gray shades and error bars represent the 95% confidence interval. (**b**) Fixation map consistency over viewing duration. (**c**) Correlations between fixation map consistency and fixation count over viewing duration. (**d**) Correlations between fixation map consistency and recognition accuracy over viewing duration. (**e**) The Edinburgh results at 2 seconds. The filled square and triangles indicate the scenes presented in Fig. 1. (**f**) Correlations between fixation count and recognition accuracy over viewing duration.

Fig. 3a illustrates that the Edinburgh fixation count linearly increased with viewing duration, as expected, and the Edinburgh fixation count at 2 seconds was not significantly different to the FIGRIM fixation count. Fig. 3b illustrates that the Edinburgh fixation map consistency was relatively stable across viewing duration, and the Edinburgh fixation map consistency at 2 seconds was not significantly different from the FIGRIM fixation map consistency. These results suggest that despite the many differences, the calculation of fixation counts and fixation map consistency was robust across the two datasets.

Fig. 3c illustrates that the correlation between fixation count and fixation map consistency in the Edinburgh dataset were not significantly different across the different viewing durations analyzed. In the FIGRIM dataset, the correlation between fixation count and fixation map consistency was significantly negative, $r(628) = -0.15$, 95% CI [-0.22, -0.07], $p < 0.001$, but within the 95% confidence interval of the Edinburgh correlation at 2 seconds, $r(130) = 0.00$, 95% CI [-0.17, 0.17], $p = 0.99$. These results suggest that fixation count and fixation map consistency were not positively correlated and thus may reflect different intrinsic properties of a scene, which would differently contribute to scene encoding.

Fig. 3d illustrates that the correlation values between fixation map consistency and scene memory were significantly positive at 2 seconds and afterward in the Edinburgh dataset. Especially, the maximum correlation was found at 2 seconds, $r(130) = 0.34$, 95% CI [0.18, 0.48], $p < 0.001$ (Fig. 3e). The FIGRIM correlation value, $r(628) = 0.22$, 95% CI [0.14, 0.29], $p < 0.001$ (the left panel in Fig. 2b), was within the 95% confidence interval of the Edinburgh correlation at 2 seconds. These results suggest that the fixation map consistency calculated with 2 seconds of fixation data is reliably associated with scene memory.

Fig. 3f illustrates that the correlation values between fixation count and scene memory were positive over time in the Edinburgh dataset; the correlation value at 1 second was significantly positive, $r(130) = 0.25$, 95% CI [0.08, 0.41], $p = 0.004$. Also, the FIGRIM correlation value, $r(628) = 0.19$, 95% CI [0.11, 0.26], $p < 0.001$ (the right panel in Fig. 2b) was within the 95% confidence interval of the Edinburgh correlation at 2 seconds. The comparable range of correlation values suggests that fixation count and scene memory are positively and weakly correlated and that a large number of scenes (e.g., 630 scenes in the FIGRIM dataset) is required to detect a significant relationship.

Examination of the potential mechanisms for the relationship between fixation map consistency and scene memory

In both the Edinburgh and the FIGRIM datasets, fixation map consistency and scene memory were significantly and positively correlated. In order to investigate the potential mechanisms that drive the relationship between fixation map consistency and scene memory, we examined the effects of scene category (as provided in the FIGRIM dataset, including mountain, playground, cockpit, kitchen) and scene semantics (e.g., whether people or faces were presented in the scene) on scene memory, which are associated with scene memorability (Bylinskii et al., 2015; Isola et al., 2011).

To examine the effects of scene category, we conducted a scene-level linear regression analysis on the FIGRIM dataset, in which the dependent variable was the scene memorability scores from the AMT participants, and the predictors were z-scored fixation map consistency and fixation count from the lab participants, scene category, the interaction of fixation map consistency and scene category, and the interaction of fixation count and scene category, i.e., scene memorability ~ scene category * (z-scored fixation map consistency + z-scored fixation count). This model explained ($df = 567$) explained 16.8% of the variance (adjusted R^2) and showed a significant effect of scene category, $F(20,567) = 4.24$, $p < 0.001$, consistent with the FIGRIM result that some scene categories, such as amusement park and playground, were more memorable than others, such as cockpit and highway (Bylinskii et al., 2015). But, this model showed nonsignificant interactions between scene category and fixation map consistency, $F(20,567) = 0.43$, $p = 0.97$, and between scene category and fixation count, $F(20,567) = 1.03$, $p = 0.42$, suggesting that scene category did not significantly affect the relationship between fixation map consistency and scene memory. Fig. 4a illustrates the relationship between fixation map consistency and scene memory for each scene category type. In 20 out of 21 scene categories in the FIGRIM dataset, fixation map consistency and scene memory were positively associated, with the correlation values ranging from -0.01 to 0.50 (M=0.24, SD=0.15).

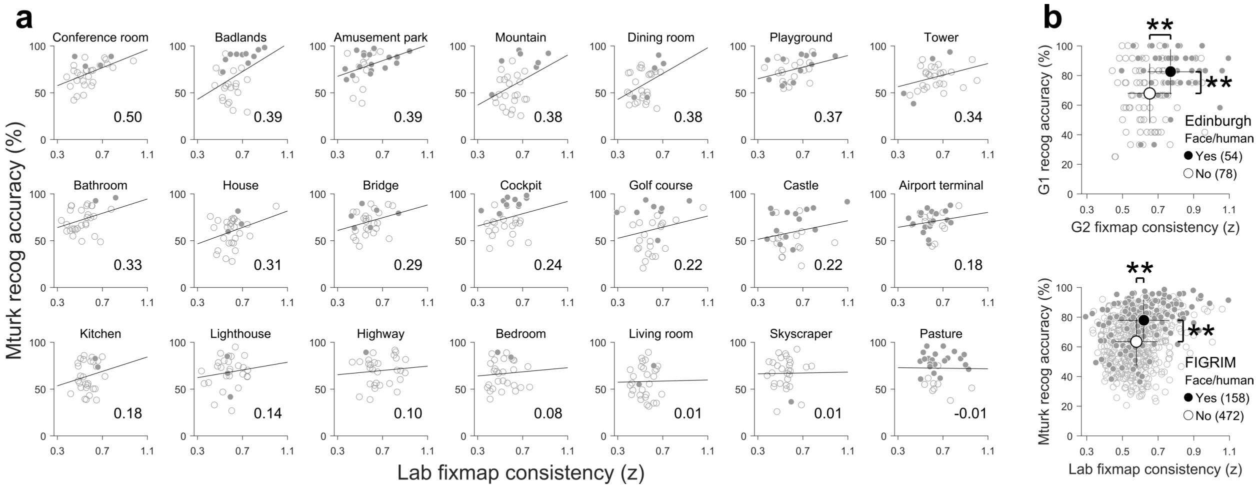

Fig. 4. Relationships between fixation map consistency, scene category, scene semantics, and scene memory. (**a**) The effect of scene category in the FIGRIM dataset. The name of each category is shown at the top of each panel. A filled circle represents a scene with face/human, and an open circle represents a scene without face/human. The line represents a linear regression. (**b**) The effect of scene semantics in the Edinburgh (top) and FIGRIM (bottom) datasets. The filled and open large circles indicate the average values across the scenes with and without face/human, respectively. The error bars indicate S.D. ** $p < 0.001$.

While scene category did not interact with memory accuracy and fixation map consistency, previous research has shown that specific scene semantics/features, such as faces and symbols, play an important role in guiding attention deployment (Cerf et al., 2009; Henderson, 2003; Henderson & Hayes, 2017, 2018; Wu, Wick, & Pomplun, 2014; Xu et al., 2014) and in forming scene memory (Isola et al., 2011). To examine the effects of scene semantics, we manually categorized the scenes in the Edinburgh and FIGRIM datasets based on whether the scene had faces (of humans, animals, and objects that have facial features in a coherent manner like a giant face on the building or Thomas the train) vs. not and compared fixation map consistency and scene memory of those scene groups (Fig. 4b). We focused on face/human because this feature has been shown to strongly attract attention (Cerf et al., 2009; Xu et al., 2014) and to be positively associated with scene memory (Isola et al., 2011). We conducted one-way ANOVA tests on fixation map consistency and scene memory in both datasets and found that the scenes with a face showed significantly higher fixation map consistency (Edinburgh: $F(1,130) = 27.5$, $p < 0.001$; FIGRIM: $F(1,628) = 12.7$, $p < 0.001$) and scene memory (Edinburgh: $F(1,130) = 20.4$, $p < 0.001$; FIGRIM: $F(1,628) = 100.6$, $p < 0.001$) than the scenes without a face. These results support the idea that scene semantics play an important role in scene memory. However, we also found no differences in the relationship between fixation map consistency and scene memory by different scene. More specifically, we found that the correlation between fixation map consistency and scene memory was similar across semantic categories in both Edinburgh (with-face/human: $r(52) = 0.06$, 95% CI [-0.21, 0.32], $p = 0.67$; without-face/human: $r(76) = 0.14$, 95% CI [-0.09, 0.35], $p = 0.23$) and FIGRIM (with-face/human: $r(156) = 0.37$, 95% CI [0.23, 0.50], $p < 0.001$; without-face/human: $r(470) = 0.12$, 95% CI [0.02, 0.20], $p = 0.01$) datasets. This suggests that scene semantics cannot fully explain the relationship between fixation map consistency and scene memory and that the relationship between fixation map consistency and scene memory still holds. Taken together, these results imply that fixation map consistency discloses an intrinsic scene property that is distinct from elaboration, scene category, and scene semantics and that is robustly associated with encoding a scene.

A potential origin of fixation map consistency

Fixation map consistency is cleary tied to scene memory, but it is unclear why and how fixation map consistency is associated with scene memory? To answer this question, it is important to understand what kind of information fixation map consistency can disclose about a scene. Previous research (Wilming et al., 2011) has only provided descriptive accounts on this topic. For example, Wilming et al., 2011 examined how an urban scene could produce higher fixation map consistency than a nature scene by relating fixation map consistency to scene semantics (e.g., whether a scene has specific visual features, such as people and man-made objects). Consistently, we found that scene semantics was positively associated with fixation map consistency (Fig. 4b). However, the mere existence of semantic features could not fully explain why higher fixation map consistency was related to higher scene memorability. In an attempt to understand this relationship, we sought to explore how a scene can produce more or less consistent fixation maps by creating a simple computational model.

Our model is inspired by the finding that semantic features have different strengths in attracting attention and fixations (Xu et al., 2014); for example, the strength of faces in attracting overt attention was about twice that of man-made objects designed to be watched (e.g., monitors). In the model (see Methods for details), the strength of fixation attraction of each pixel was quantified using a probability density function (PDF) of fixations, which was then used to simulate the fixations of virtual participants. The fixation PDF consisted of uniformly-distributed noise and a Gaussian signal at the center. We then varied the amplitude of the signal relative to noise to examine the relationship between signal-to-noise ratio (SNR) of a scene and its fixation map consistency. Fig. 5a shows the simplest case, for example, a blank screen. In this case, we assumed that no particular pixel can attract more fixations than any other pixels and thus used a uniform distribution to model the ambient noise (the top panel). Fig. 5b shows the case when there is a weak signal (the red curve), for example, a small dot present at the center of a blank screen. In this case, the peak amplitude of the Gaussian signal was set to match the amplitude of uniform noise (the dashed horizontal line), thus the SNR was defined to be one. Fig. 5c shows the cases when there are stronger signals, such as a face, present at the center (SNRs of 5 and 10, respectively); the peak amplitude of the Gaussian signal was set to 5 and 10 times of the amplitude of uniform noise, respectively. All of the fixation PDFs were scaled so that the area-under-curve (the shaded areas in the top panels) was held constant across the different SNR conditions.

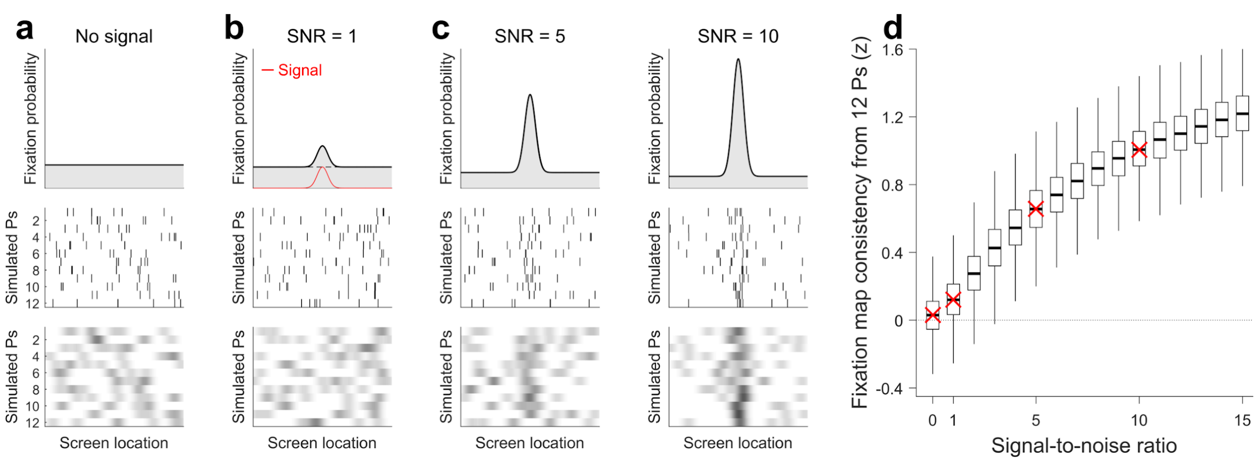

Fig. 5. A potential origin of fixation map consistency. (**a**) Simulated fixations when there is no signal, for example, a blank screen. In all panels, the horizontal axis specifies the location of 800 pixels in a 1-dimensional screen. In the top panel, the probability density function (PDF) of 1-D fixation locations is shown (the horizontal black line), in which a uniform distribution was used to model ambient noise. The shaded area was used to calculate the area-under-curve (AUC). In the middle panel, each row represents a virtual participant, and each vertical bar represents a simulated fixation on 1-D screen. Seven random fixations were generated based on the PDF for

each of the twelve virtual participants. The bottom panel shows the smoothed 1-D fixation maps of the twelve virtual participants in the same order. (**b**) Simulated fixations when a weak signal is present (i.e., signal-to-noise ratio (SNR) of 1). The PDF of simulated fixations was generated by adding a uniform noise (the black dashed horizontal line) and a Gaussian signal (the red solid curve) and the peak amplitude that equals that of the uniform noise. Then, the PDF was scaled so that its AUC is equal to that of the PDF in (**a**). (**c**) Simulated fixations when stronger signals are present (i.e., SNRs of 5 and 10, respectively). The PDFs were generated and scaled in the same manner as in (**b**). As a result, the AUCs in all the top four panels are the same, and the fixations were simulated using the scaled PDFs. (**d**) Fixation map consistency as a function of SNR. This box plot illustrates the distribution of fixation map consistency from 2000 simulated sets of twelve virtual participants in each SNR condition. The upper and lower horizontal edges of the rectangles specify the first and third quartile, the middle thick lines specify the median, and the red Xs represent the exemplary sets shown in the panels **a-c**.

We varied SNR from 0 to 15 in increments of 1 and simulated 2000 sets of twelve virtual participants for each SNR by randomly generating seven fixations for each participant (e.g., the middle panels in Figs. 5a-c) using the PDF associated with that SNR. The number of fixations was set to seven because we assumed the viewing duration of 2 s. Next, we generated individual fixation maps by smoothing the fixations using the same-sized Gaussian kernel as we did in the Edinburgh dataset. The bottom panels in Figs. 5a-c show that the individual fixation maps become more similar to each other as SNR increases. This observation was confirmed by illustrating the relationship between SNR and fixation map consistency (Fig. 5d), which was calculated for each set of twelve virtual participants using the same procedure (Methods). When the signal was not present or weak (e.g., SNR of 1), fixation map consistency was not significantly different from the chance level of 0. When the signal was strong, the signal could attract fixations consistently across viewers and thus produced a high level of fixation map consistency. Together, these results suggest that the higher the strength of (localized) signal in a scene in attracting fixations, the higher its fixation map consistency.

Additionally, the computational approach allowed us to examine the number of viewers necessary to obtain reliable fixation map consistency. To do so, we repeated the simulation with different numbers (6, 24, and 48) of virtual participants in a set and examined whether and how fixation map consistency is affected (Fig. 6). We observed that a low number of viewers (the filled circles) resulted in lower and more variable estimates of fixation map consistency because the averaged fixation map that is compared against the leave-one-out individual fixation map becomes less stable with less viewers. In contrast, a higher number of viewers (the filled triangles and rectangles) resulted in higher and less variable estimates of fixation map consistency. However, those estimates and SD all stayed within the shaded area (± SD of 12 virtual participants). These results suggest that 12-16 participants, in accordance with the Edinburgh and FIGRIM datasets, can produce sufficiently stable fixation map consistency, consistent with Wilming and colleagues' suggestion of 20 participants for practical applications (Wilming et al., 2011).

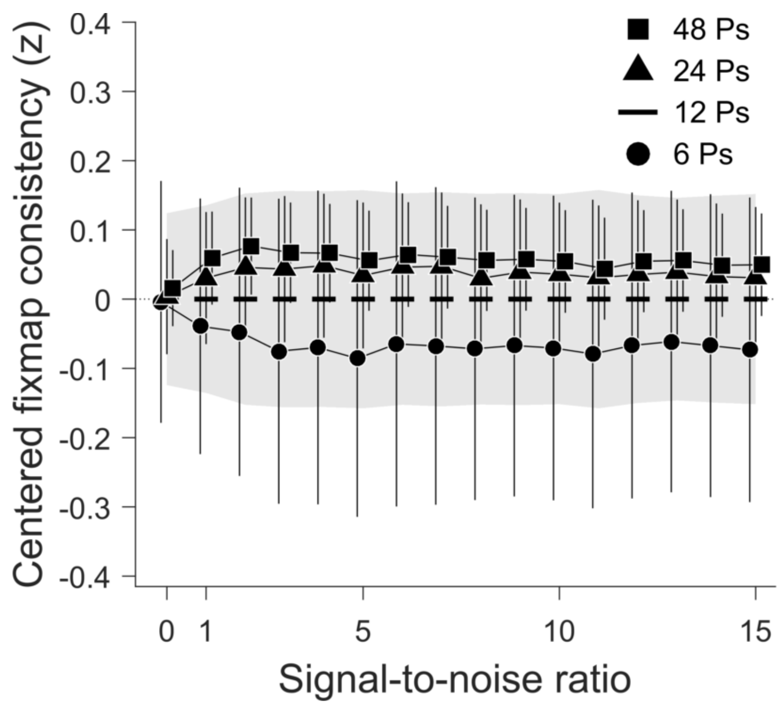

Fig. 6. Stability and systematic bias of fixation map consistency. Fixation map consistency values were centered using the mean fixation map consistency values in each SNR condition with twelve virtual participants to highlight the systematic bias driven by the number of participants. The shaded area specifies ± standard deviation (SD) of the twelve-participant condition. The circle, triangle, and square symbols indicate the number of 6, 24, and 48 virtual participants in a set, respectively, and the error bars specify the SDs.

Discussion

In this study, we used two different datasets, the Edinburgh and FIGRIM datasets, to examine whether and how fixation map consistency uniquely contributes to scene memory that is different from fixation count. Although the Edinburgh and FIGRIM datasets were different in many aspects, we found that fixation map consistency, fixation count, and their relationships to scene memory were similar between these datasets (Fig. 3). In particular, we found that fixation map consistency measured from one group of participants was significantly and positively associated with scene memory measured from a different group of participants in both datasets (Fig. 2), consistent with previous research (Khosla et al., 2015; Mancas & Le Meur, 2013). Moreover, the observed correlation values between fixation map consistency and scene memorability from the Edinburgh and FIGRIM datasets (0.25 [0.08, 0.40] and 0.22 [0.14, 0.29], respectively) were very similar to the values reported by Khosla et al. (0.24) from the Fixation Flickr dataset (Judd, Ehinger, Durand, & Torralba, 2009). Together, these findings suggest a robust positive association between fixation map consistency and scene memory.

We also found that the correlation between fixation count and scene memory was significantly positive (0.19 [0.11, 0.26]) in the FIGRIM dataset and nonsignificantly positive (0.14 [−0.04, 0.30]) in the Edinburgh dataset, consistent with previous research (Choe et al., 2017; Loftus, 1972; Tatler & Tatler, 2013). Moreover, we found that fixation count was not significantly positively correlated with fixation map consistency (Figs. 2b, 2d, and 3c), suggesting that these measures represent different information and would differently contribute differentially to scene memory. As a result, the full linear regression models (i.e., Models Ef and Ff) that utilized both fixation counts and fixation map consistency better predicted scene memorability than using either measure in isolation (Figs. 2c and 2f), although the extent of improvement was different between the two datasets. For practitioners who want to use eye-tracking measures to predict scene memory, we suggest that two seconds of eye-tracking (Fig. 3) with 16 passive-viewing participants (Fig. 6) would produce sufficiently useful population-level fixation count and fixation map consistency data. These eye-tracking measures may provide additional information to the computer vision-based scene memorability models (Bylinskii et al., 2015; Khosla et al., 2015) and enhance scene memorability prediction.

Why and how is fixation map consistency associated with scene memory? Although we found that scenes with particular semantic features (e.g., human faces) showed higher fixation map consistency and higher scene memory than scenes without these semantic features (Fig. 4b), the mere existence of semantic features could not fully explain why the higher fixation map consistency was related to the higher scene memorability. These results suggested that fixation map consistency discloses an intrinsic scene property about a scene that is distinct from elaborative viewing and scene category, but is robustly associated with scene memory. Then, we sought to understand how a scene can produce more or less consistent fixation maps by creating a simple computational model. Inspired by the finding that semantic features have different strengths in attracting attention and fixations (Xu et al., 2014), we assumed that some regions in a scene will attract more fixations (i.e., signal) than other regions (i.e., ambient noise) and varied the relative strength of signal to noise (i.e., SNR). Our model showed that the higher a scene's SNR, the higher its fixation map consistency (Fig. 5), suggesting that scenes with features/regions that can strongly attract fixations (i.e., higher SNR) would produce more consistent fixation maps across viewers and thus would be remembered better. For example, a scene with two people having funny poses (e.g., Fig. 1a) carries strong signal and thus higher SNR. This will produce highly consistent fixation maps and better subsequent memory for the scene.

There are two important implications of the SNR-based fixation model. First, the SNR of a scene, as measured with fixation map consistency, is an intrinsic property of a scene that has cognitive effects, like memorability (Khosla et al., 2015; Mancas & Le Meur, 2013). Our findings suggest that scenes with high SNR produce greater fixation map consistency and are more memorable. In a similar manner, videos that produced more consistent eye movement patterns were rated as more preferred (Christoforou, Christou-Champi, Constantinidou, & Theodorou, 2015) and might also be more memorable. Second, the same underlying signal may result in very different patterns of fixations depending on SNR (i.e., a low-SNR scene will produce more varied individual fixation maps; a high-SNR scene will produce highly consistent individual fixation maps that are similar to the underlying signal map). This suggests that considering the fixation attraction strength of visual features in a scene is critical in predicting human fixations.

There are at least three limitations of this study. First, although fixation count and fixation map consistency were associated with scene memory, the overall predictive power of these eye-tracking measures was low, as indicated by the explained variance of the full models (Edinburgh: 24.5%, FIGRIM: 18.0%). Second, this study 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 count and saccade rate across viewers has been less studied. Third, our computational model was far too simple, as fixations were randomly generated. Future research is necessary to make the model more realistic by explicating the relationships between fixation map consistency and the bottom-up (i.e., scene-specific) and top-down (e.g., instructions, viewing tasks, and past experience) factors that affect gaze control (Ballard & Hayhoe, 2009; Henderson, 2007, 2011, 2017; Kardan et al., 2016). Such effort will lead to a precise understanding of what fixation map consistency can tell us about a scene.

Conclusions

By examining two different eye-tracking datasets, we found that the higher the fixation map consistency of a scene, the higher its memorability is. Fixation map consistency and fixation count were not significantly correlated, suggesting that fixation map consistency encodes distinct information and differently contributes to scene memory from fixation count. Fixation map consistency was a stronger predictor of scene memory than fixation count, but, using both eye-tracking measures may better predict scene memorability than using either measure in isolation. For practitioners who want to use eye-tracking measures to predict scene memory, we suggest that 2 seconds of eye-tracking with 16 passive-viewing participants would produce sufficiently useful population-level fixation count and fixation map consistency. To provide a mechanistic explanation for how a scene can produce more or less consistent fixation maps across viewers, we created a simple computational model by assuming some regions in a scene will attract more fixations (i.e., signal) than other regions (i.e., ambient noise) and then varied the amplitude of the signal relative to noise to examine the relationship between signal-to-noise ratio (SNR) of a scene and its fixation map consistency. Our model showed that the higher a scene's SNR, the higher its fixation map consistency, suggesting that fixation map consistency reflects the relative strength of signal (for capturing attention) in a scene relative to its ambient noise, which is an intrinsic property of a scene that can affect human vision and memory.

Author Contributions

J. M. Henderson conceived of and designed the Edinburgh experiment and provided the data that were collected under his supervision. K. W. Choe and M. G. Berman developed the study concept. M. Lyu and K. W. Choe analyzed the data and drafted the manuscript. O. Kardan, H. P. Kotabe, J. M. Henderson, and M. G. Berman provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments

This work was 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. Berman, the National Science Foundation to M. G. Berman (BCS-1632445), and the National Eye Institute of the National Institutes of Health to J. M. Henderson (R01EY027792).

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.

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.

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. doi: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. doi: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*, Vol. 6. https://doi.org/10.3389/fpsyg.2015.00579

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. doi: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. doi:10.1167/16.11.13.

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*, Vol. 16, pp. 219–222. https://doi.org/10.1111/j.1467-8721.2007.00507.x

Henderson, J. M. (2011). Eye movements and scene perception. *Oxford Handbooks Online*. https://doi.org/10.1093/oxfordhb/9780199539789.013.0033

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. doi: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? *CVPR 2011*. https://doi.org/10.1109/cvpr.2011.5995721

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. *2009 IEEE 12th International Conference on Computer Vision*. https://doi.org/10.1109/iccv.2009.5459462

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.

Kardan, O., Shneidman, L., Krogh-Jespersen, S., Gaskins, S., Berman, M. G., & Woodward, A. (2017). Cultural and Developmental Influences on Overt Visual Attention to Videos. *Scientific Reports*, *7*(1), 11264.

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and Predicting Image Memorability at a Large Scale. *2015 IEEE International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/iccv.2015.275

Loftus, G. R. (1972). Eye fixations and recognition memory for pictures. *Cognitive Psychology*, Vol. 3, pp. 525–551. https://doi.org/10.1016/0010-0285(72)90021-7

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. doi:10.1167/14.14.9.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics*, Vol. 2, pp. 547–552. https://doi.org/10.3758/bf03210264

Mancas, M., & Le Meur, O. (2013). Memorability of natural scenes: The role of attention. *2013 IEEE International Conference on Image Processing*. https://doi.org/10.1109/icip.2013.6738041

Masciocchi, C. M., Mihalas, S., Parkhurst, D., & Niebur, E. (2009). Everyone knows what is interesting: salient locations which should be fixated. *Journal of Vision*, *9*(11), 25. doi:10.1167/9.11.25.

McCamy, M. B., Otero-Millan, J., Di Stasi, L. L., Macknik, S. L., & Martinez-Conde, S. (2014). Highly informative natural scene regions increase microsaccade production during visual scanning. *The Journal of Neuroscience* , *34*(8), 2956–2966.

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. doi: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*, Vol. 22, pp. 975–995. https://doi.org/10.1080/13506285.2014.941433

Onat, S., Açık, A., Schumann, F., & König, P. (2014). The contributions of image content and behavioral relevancy to overt attention. *PloS One*, *9*(4), e93254.

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

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. doi: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.

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.

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: A Journal of the Psychonomic Society, Inc*, *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.

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

Supplemental Material for “Scenes that produce more consistent fixation maps are more memorable” by Lyu, Choe, Kardan, Kotabe, Henderson, & Berman.

Supplemental Method: Edinburgh dataset

Two groups of 36 undergraduate students (Group 1 and Group 2) from the University of Edinburgh participated in the experiment. All 72 participants had 20/20 corrected or uncorrected vision, were naive to the purposes of the experiment and provided informed consent as administered by the Institutional Review Board of the University of Edinburgh.

Participants sat 90 cm away from a 21-inch CRT monitor and placed their head on a chin and forehead rest. The scenes were displayed fullscreen in their native resolution and subtended 25.8° x 19.4° in visual angle. Eye movements were recorded from the right eye, although viewing was binocular, via an SR Research (Ottawa, Ontario, Canada) Eyelink 1000 eye tracker with a sampling rate of 1000 Hz. The experiment was controlled with SR Research Experiment Builder software. The eye tracker was calibrated using a built-in nine-point calibration routine. The calibration was not accepted until the average error was less than 0.49° and the maximum error was less than 0.99°.

Both Group 1 (G1) and Group 2 (G2) participants performed the first phase (Encoding phase) of the experiment in which the viewing task was manipulated (i.e., visual search task, memorization task or aesthetic preference task). Shortly after the Encoding phase only G1 participants engaged in the second phase testing scene recognition (Test phase). So, the fixation patterns during memorization were obtained from G1 and G2, and the recognition accuracy from memorization encoding was obtained from G1 participants only. In the Encoding phase, 135 full-color (32 bit) 800 x 600-pixel photographs of real-world, indoor and outdoor scenes were presented. The 135 scenes were split into three blocks of 45, and the scene split was the same across participants. During each block, participants were instructed to perform one of three tasks on the scenes presented for 8 s while their eye movements were recorded: (1) memorize the scene for a subsequent old/new recognition test (but no response was required during the encoding period), (2) search for an object, or (3) make an aesthetic preference judgment. The task assignment and order of each block were determined by a dual-Latin square design and counterbalanced across participants. Within each task block, the scenes were presented in random order. The participants completed all three blocks. In this paper, we focus on the intentional scene memorization encoding task and resulting memory performance. For more details and results from the visual search and preference judgment tasks, see Choe et al. (2017).

After completing all three task blocks and a short break, G1 participants engaged in the Test phase, i.e., the scene recognition task. Before the task, participants were informed that their memory would be tested for all of the scenes they had previously encountered, not just the scenes they had been instructed to remember in the memorization block. In each trial, a scene was shown for 3 s, 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'. In total, 66 of those scenes were 'old', i.e., seen in the encoding phase, and the other 66 scenes were 'altered' stimuli, and the remaining 3 scenes were not used in the Test phase. In addition, 22 new scenes were never before seen stimuli. In this paper, we present the results of the 132 scenes that were used in both the Encoding and Test phases. The scene images are available from the author J.M.H. upon request. A total of 154 scenes, consisting of seven categories that contained 22 scenes – old & with memory encoding, old & search encoding, old & preference encoding, altered & memory encoding, altered & search encoding, altered & preference encoding, and new – were used in the recognition task. The recognition accuracy used in this study was based solely on the old & memory and altered & memory trials only (44 trials per participant).