Supplement to: Objects do not predict fixations better than early saliency; A re-analysis of Einhäuser *et al.*'s data

Ali Borji, Dicky. N. Sihite, Laurent Itti

June 1, 2013

1 Methods

1.1 Scoring Methods

We use three prevalent scoring methods to test the object-map hypothesis, since validity of this model boils down to fair model comparison with bottom-up saliency models. Please see Appendix B for explanation of state-of-the-art saliency models used here. We report results using the previously proposed Normalized Scanpath Saliency (NSS) (Peters et al., 2005; Parkhurst et al., 2002), Correlation Coefficient (CC), and Shuffled AUC score (AUC Type-3) (Parkhurst & Niebur, 2003; Tatler, 2007). Our emphasis is more on the Shuffled AUC score as it is the only score that tackles center-bias at the data level. Please see Appendix C for detailed explanation of scoring methods.

1.2 Object-based Models

Here we present a formal definition of the object-based model by Einhäuser et al. (2008). Let B_j be a binary map with "1" at the location of the annotated object j (out of $N = 981$ unique labels) and zeros elsewhere. Assume K subjects have performed the task. As defined by Einhäuser et al. (2008), the predicted object map for each image is the summation of annotated object maps weighted by their recall frequency:

$$\text{object\_map} = \frac{1}{KN} \sum_{i=1}^{K} \sum_{j=1}^{N} w_{ij} B_j \tag{1}$$

where w_{ij} is 1 if object j is remembered by subject i and 0, otherwise (i.e., weighted by the object recall frequency). In addition to Einhäuser's original model, we investigate the fixation prediction power of three additional variants of the object map model: 1) a map made of first remembered objects by subjects, 2) a map made of the most remembered object by all subjects, and 3) a map built from the object that has been most remembered first. Please see Appendix D for a detailed description of object-map models.

We compare four object-based models (Einhäuser's original and our three variants, as representatives of the object-based hypothesis) to eleven bottom-up saliency models (as representatives of the saliency hypothesis). We employ three types of ITTI model maps: two different versions of the Itti *et al.* model, ITTI98 and ITTI, which correspond to different normalization schemes. We also use the exact saliency maps of Einhäuser et al. (2008) (denoted here as ITTI*), to make our results directly comparable.

2 Analysis Results

2.1 Scoring Metrics and Center-bias

Fig. S1 shows ROC plots and AUC values for all models. As it can be seen, the original object-map model (See case 1 in Appendix D) scores significantly below many saliency models, using all three types of AUC

scores (paired t-test, $\alpha < 0.05/n = 0.0045$ with Bonferroni correction to the number of compared non-trivial models, thus $n = 11$ comparisons).

Results of our *first analysis* (See main text) shows that the object-map model is significantly better than the ITTI* model using AUC Type-1 (in alignment with Einhäuser et al. (2008)) and AUC Type-2 — two metrics which do not discount center bias. However, using AUC Type-3 which discounts the center-bias in data, there is not a significant difference between saliency and objects' ability to predict human gaze anymore.

Please note that Einhäuser et al. (2008) superimposed fixations from another randomly chosen image to measure the amount of center-bias (called random assignment procedure). They showed, both for ITTI* saliency model and their object-map model, that gaze patterns over one image could significantly be predicted by the model map from another, randomly chosen, image. This showed that a common bias (center bias) existed in all images. However, gaze to model agreement scores were significantly lower when using randomly chosen maps compared to using each image's true corresponding map. Thus, center bias alone could not explain the entire human data. In their results, the saliency model scored 0.578 ± 0.076, of which they argued a score of 0.529 ± 0.057 may be due to center bias alone. Similarly, their object-based map scored 0.651 ± 0.106, of which 0.598 ± 0.107 may be due to center bias. Interestingly, while Einhäuser et al. (2008) were aware of center-bias in fixation data and showed that the prediction accuracy of their proposed object map exceeds center-bias chance level, they did not directly compare saliency and object-based maps after discounting center-bias (i.e., comparing residuals).

The object-map model is not significantly better than the ITTI model with the AUC Type-1, although it is significantly better using the other two types of AUC scores (due to sparseness of this model). The original, older ITTI98 model scores significantly higher than the object-map model using three types of AUC scores. ITTI98 also scores better than ITTI on this particular comparison to human gaze data, thanks to the smoother maps of the ITTI98 model. Performances of the object-map and ITTI* models are listed in Table 1. Combining the object map to the ITTI* model (adding maps) scores 0.576 ± 0.069 (AUC Type-3) which is significantly above the ITTI* alone (paired t-test; $p = 2.23 \times 10^{-6} < 0.05$) as well as the object-map alone (paired t-test; $p = 0.0048 < 0.05$). Thus, object information helps fixation prediction (i.e., provides an independent source of information than what is conveyed by the ITTI* model), but alone does not perform significantly above saliency.

Please see Appendix D for other variants of the object-map model and their performances. The accuracy of the weighted object map (i.e., Einhäuser *et al.*'s original, case 1) exceeds the other variants of the object-map model as well as the unweighted object-map[1].

Two other scores which have been frequently used in the past are CC (correlation coefficient between smoothed human fixation map and a model's prediction map) and NSS (average activity at human fixations in a normalized prediction map). These scores are contaminated by center-bias (Borji et al. (2012)). A central Gaussian (as well as the mean-eye position (MEP) map) using these two scores outperforms almost all state-of-the-art saliency models (similar to AUC Types 1 and 2). Thus, we recommend not using these scores in the future because they are overwhelmed by center bias. However, here for the sake of completeness, we show the results using CC and NSS scores (See Fig. S6; Appendix C).

2.2 Selected Models

In our *second analysis* (See main text) we used different saliency models as representatives of the saliency hypothesis[2]. The human inter-observer model, which is a smoothed map built from fixations of other subjects over the same image, achieves the highest score (0.803 ± 0.111; mean $\pm$ standard deviation; AUC Type-3). The Uniform random model scores around 0.5 using all three types of AUC. Note that although the Normal

[1]One may argue that "object maps" of Einhäuser et al. (2008) are the crudest possible "object-based" models one could possibly think of. Indeed, more sophisticated object-based models may exist and may even work better than all considered saliency models here. Our results here directly address their claim based on their model, and hold until the proof of the contrary and discovery of better object-based models.

[2]Note that saliency is not a unique concept but all saliency models are based on bottom-up image outliers, which is in contrast to object-based factors. In this study, we chose those models that only use purely bottom-up features.

Table 1: Accuracy of the object-map vs. the ITTI* model (mean ± standard deviation) along with the test of statistical significance using paired t-test over 93 images.

Score	*Object-map (case 1)*	*ITTI**
AUC Type-1	$0.6480^{\dagger} \pm 0.107$	0.5841 ± 0.076
AUC Type-2	0.6417 ± 0.107	0.5762 ± 0.010
AUC Type-3	$0.5590 \pm 0.080^{n.s.}$	0.5467 ± 0.060
NSS	0.6120 ± 0.471	0.4142 ± 0.421
CC	0.1520 ± 0.114	0.1026 ± 0.103

$^{n.s.}$ Not significant at $p < 0.05/n = 0.0045$ with Bonferroni correction (p = 0.2343, See Fi.g S2).

* Saliency maps used in Einhäuser et al. (2008).

† This is close to 0.651 ± 0.106 reported in Einhäuser et al. (2008).

random model, a central Gaussian with sigma of 5.8° × 4.4° (See Fig.1 main text), overcomes almost all other models using AUC Types 1 and 2, it is discounted using the AUC Type-3, scoring about 0.5. Every difference between the object-map model and other models is significant (except the ITTI* model) using the AUC Type-3, with the object-map model being significantly above the very sparse ITTI model but being below the smoother ITTI98 and all other tested models.

Using AUC Type-1 like Einhäuser et al. (2008), the ITTI* model used by Einhäuser *et al.* scores 0.578 ± 0.076. Note how the prediction power of both the very sparse ITTI model (AUC Type-1 scores 0.633 ± 0.040) and the smoother ITTI98 model (AUC Type-1 scores 0.689 ± 0.093) are higher than the score obtained by Einhäuser *et al.* with the ITTI* model (which seems to involve some custom settings and map thresholding over the published ITTI98 model). Since models have different parameters and versions, replicating exact values is very difficult. This is another reason for not establishing conclusions based on just one model[3]. Using AUC Type-1, the prediction power of the ITTI98 is 0.689 ± 0.093 and the maximum among models belongs to the GBVS model (0.786 ± 0.084). MEP (Mean Eye Position over all images) scores 0.792 ± 0.067 and stands above all models due to center-bias in the eye data. Performances of these four models (ITTI98, AIM, AWS, and MEP) using AUC Type-3, in order are: 0.590 ± 0.079, 0.637 ± 0.077, 0.647 ± 0.084 and 0.491 ± 0.059. Note in particular how MEP now scores at chance level with AUC Type-3.

Since some researchers have used the Matlab implementation by D. Walther of the ITTI model (Itti et al., 1998), known as the Saliency Toolbox (STB; http://www.saliencytoolbox.net/), here we investigate the predictive power of this model. The STB model achieves 0.535 ± 0.061, 0.520 ± 0.035, and 0.509 ± 0.030 using AUC types 1 to 3, respectively (all significantly below the corresponding values of the object-map case-1; using paired t-test $\alpha = 0.05$). Overall, the original ITTI98 and ITTI models both perform significantly higher than the STB model over this dataset. From this analysis and our model benchmarking study (Borji et al., 2012), we observe lower performances for the STB model compared with either the ITTI or ITTI98 models over free-viewing datasets. The reasons might be different implementations of scale pyramids, normalization schemes, or blurring techniques in the two approaches. Thus, one should be careful when using the STB model as a representative of the ITTI98 or ITTI models for behavioral data analysis, as STB yields significantly different saliency maps compared to ITTI or ITTI98 models. We suggest using high performing saliency models.

[3]Not only here we do investigate the object-map hypothesis based on the same parameter setting used in Einhäuser et al. (2008), we also search in a larger parameter space to test the validity of this hypothesis.

References

Avraham, T., & Lindenbaum, M. (2010). Esaliency (extended saliency): Meaningful attention using stochastic image modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, *32*, 693–708.

Berg, D., Boehnke, S., Marino, R., Munoz, D., & Itti, L. (2009). Free viewing of dynamic stimuli by humans and monkeys. *Journal of Vision.*, *9*, 1–15.

Borji, A., Sihite, D. N., & Itti, L. (2012). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Trans. Image Processing.*, *22*, 55–69.

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

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*.

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2007). Predicting human gaze using low-level saliency combined with face detection. *Advances in Neural Information Processing Systems (NIPS).*, *20*, 241–248.

Ehinger, K., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modeling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition.*, *17*, 945–978.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision.*, *8*, 1–26.

Fukuchi, M., Tsuchiya, N., & Koch, C. (2010). The focus of expansion in optical flow fields acts as a strong cue for visual attention. *Journal of Vision.*, *9*.

Gao, D., Han, S., & Vasconcelos, N. (2009). Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Trans. PAMI.*, *31*.

Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., & Dosil, R. (2012a). Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing.*, *30*, 51–64.

Garcia-Diaz, A., Leboran, V., Fdez-Vidal, X. R., & Pardo, X. M. (2012b). On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of Vision.*, *12*.

Green, D., & Swets, J. (1966). *Signal Detection Theory and Psychophysics.*. New York: John Wiley.

Guo, C., & Zhang, L. (2010). A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Trans. on Image Processing.*, *19*.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Advances in Neural Information Processing Systems (NIPS).*, *19*, 545–552.

Hou, X., Harel, J., & Koch, C. (2012). Image signature: Highlighting sparse salient regions. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI).*, *34*, 194–201.

Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Hou, X., & Zhang, L. (2008). Dynamic visual attention: Searching for coding length increments. *Advances in Neural Information Processing Systems (NIPS).*, (pp. 681–688).

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition.*, *12*, 1093–1123.

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems (NIPS)*.

Itti, L., & Koch, C. (1999). A comparison of feature combination strategies for saliency-based visual attention systems. *SPIE human vision and electronic imaging IV.*, *3644*, 473–482.

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

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 1254–1259.

Jost, T., Ouerhani, N., von Wartburg, R., Mäuri, R., & Häugli, H. (2005). Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding.*, *100*.

Judd, T., Ehinger, K., & Durand, a. T. A., F. (2009). Learning to predict where humans look. In *International Conference on Computer Vision (ICCV)*.

Kienzle, W., Franz, M., Schölkopf, B., & Wichmann, F. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision.*, *9*.

Kootstra, G., & Schomaker, L. (2009). Prediction of human eye fixations using symmetry. In *Proc. Conference on Cognitive Science (CogSci09)*.

Mancas, M. (2007). *Computational Attention: Modelisation and Application to Audio and Image Processing.*. PhD. thesis.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision.*, *42*, 145–175.

Pang, D., Kimura, A., Takeuchi, T., Yamato, J., & Kashino, K. (2008). A stochastic model of selective visual attention with a dynamic bayesian network. In *ICME*.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research.*, *42*, 107–123.

Parkhurst, D., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision.*, *16*, 125–154.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, *45*, 2397–2416.

Rajashekar, U., van der Linde, I., Bovik, A., & Cormack, L. (2008). Gaffe: A gaze-attentive fixation finding engine. *IEEE Transactions on Image Processing.*, *17*, 564–573.

Shaffer, J. (1995). Multiple hypothesis testing. *Annual Review of Psych.*, *46*, 561–584.

Shore, S., Tillman, L., & Schmidt-Wulffen, S. (2004). *Stephen shore: Uncommon places: The Complete works.*. New York: Aperture.

Tatler, B. (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*, 1–17.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, *113*, 766–786.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision.*, *8*.

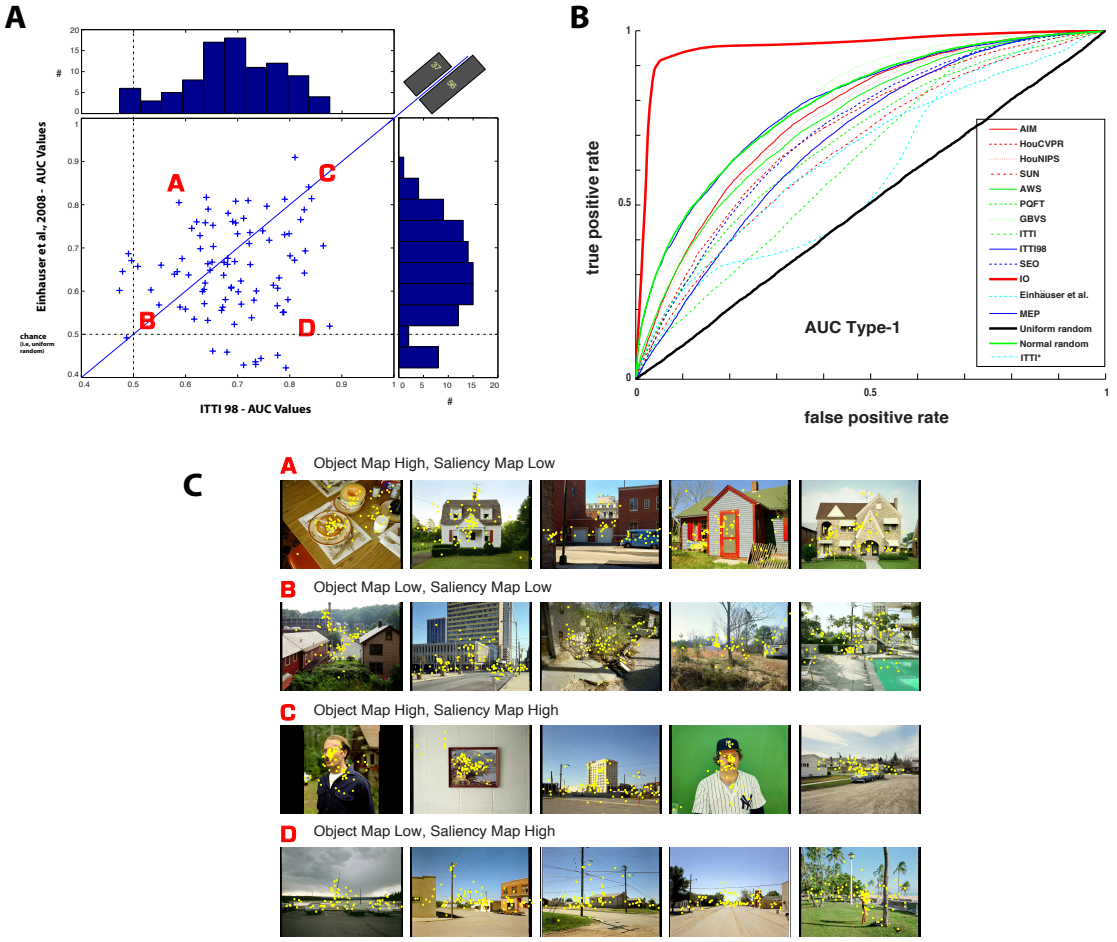

Figure S1: **A)** Area under the curve (AUC Type-1) for fixation prediction using ITTI98 model (x-axis) vs. object-map model of Einhäuser *et al.* (y-axis). Each data point corresponds to one image. Distribution of either AUCs depicted as marginals (same axes as the scatter plot). For 56 images, ITTI98 model achieves higher scores. ITTI98 model scores lower than chance (AUC = 0.5) for only 5 images (9 images for the object-map model). **B)** ROC curves of models with Type-1 calculation (over all images). Note that because AUC Type-1 does not discount center-bias in data, MEP outperforms many models. Normal random which is a central Gaussian works as well as the MEP and above all other models. Uniform random model ranks at the bottom. **C)** We investigate the model agreement by showing images at four corners of the plot in A.

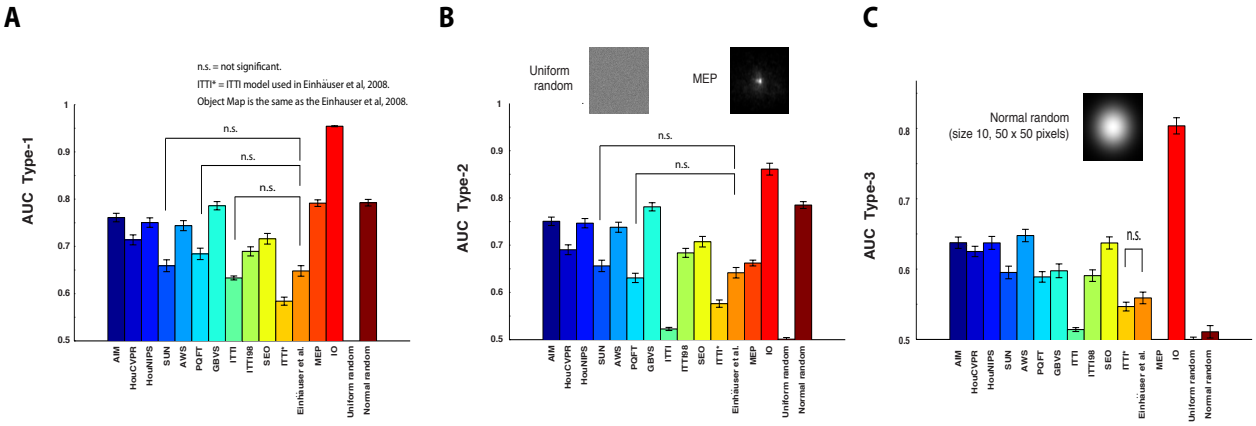

Figure S2: AUC values of three types of AUC score calculation. Error bars indicate standard error of the mean (SEM): $\sigma/\sqrt{Q}$, where σ is the standard deviation and $Q = 93$ is the number of images. Here, we only show the cases where t-test value is not significant. To account for multiple model comparisons, we correct the statistical significance using Bonferroni correction (Shaffer, 1995). Since we have 11 models, we choose α equal to $0.05/11 = 0.0045$. Here we are only interested in comparison of models with the object-map model and not ranking models over this data although some models obviously work better than the others. **A)** Using AUC Type-1, (used by Einhäuser et al., 2008) the object-map is not significantly better than any other model except the ITTI* model ((consistent with Einhäuser et al., 2008) due to center-bias). **B)** Using AUC Type-2, the object-map model is significantly better than ITTI (due to sparseness of its maps) and ITTI* models. **C)** With AUC Type-3 (shuffled AUC score), which discounts center-bias in eye data, every difference between the object-map and other models is statistically significant except the ITTI* model. AWS scores the best. MEP and the random model score lowest about 0.5. Over all three types of AUC, the object-map model is significantly below the ITTI98. Note the big gap among models and the Inter-observer model. This supports our second analysis about choosing the right model for data analysis (i.e., dependency of conclusions to the used model). Taking ITTI model as the representative of the bottom-up attention, an object-based representation explains fixation better (only using AUC Types 1 and 2 which are contaminated by center-bias and not AUC Type 3), while with newer models or with the original ITTI98 model (with a different normalization scheme than newer ITTI model resulting in smoother maps), image-based outliers predict fixations significantly better. This result also shows the importance of appropriate tackling of center-bias in fixation data. While a simple MEP (Mean Eye Position map) or a Normal random distribution seems to explain fixations better at the first glance, with an appropriate measure (here shuffled AUC score), it is clear that such models fall short for explaining eye fixation data (specially those fixations that fall off center). To build a Gaussian of sigma size $0.28^\circ \times 0.28^\circ$ on this dataset, we used this Matlab script: *myGauss=fspecial('gaussian',50,10)* where 50 and 10 are the image size and standard deviation of the Gaussian, respectively.

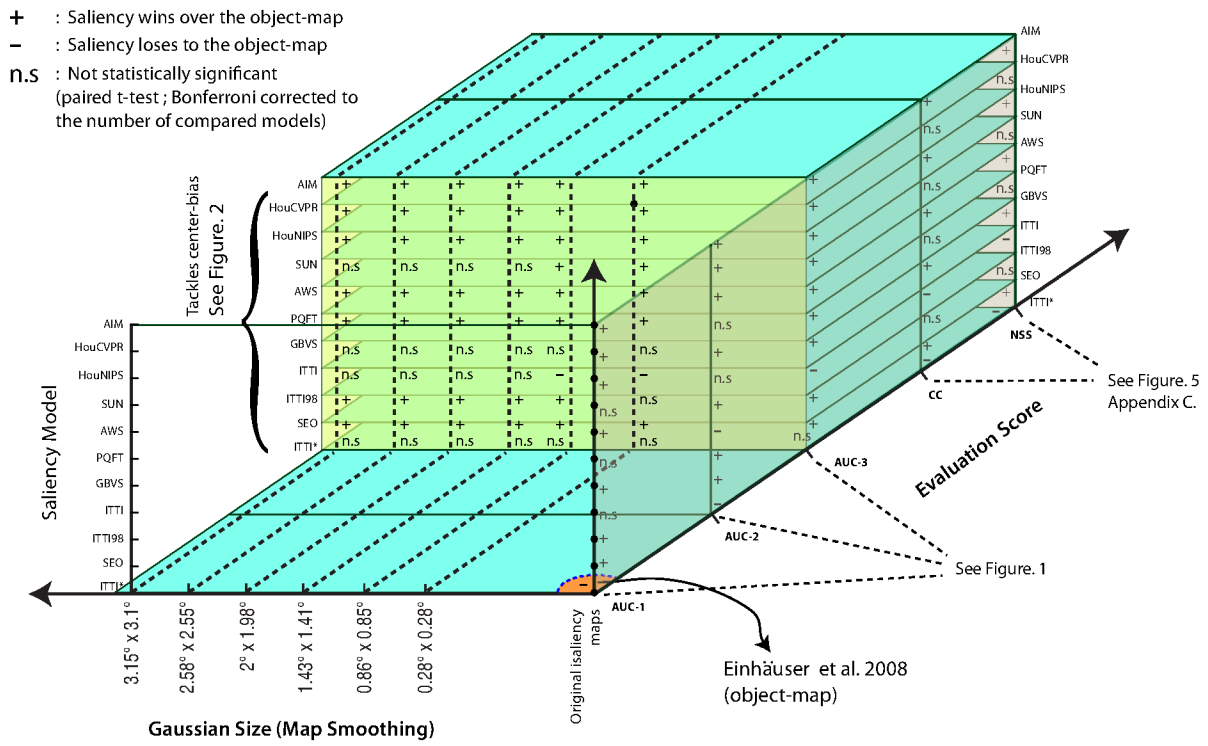

Figure S3: Summary of our results using Bonferroni-corrected paired t-test to the number of compared models (n=11; i.e., object-map vs. 11 saliency models; $\alpha = 0.0045$). We investigated accuracy of the object-map hypothesis on three dimensions: 1) Saliency model, 2) Evaluation score, and 3) Gaussian size used for map smoothing. This hypercube includes a much wider parameter space considered in Einhäuser et al. (2008) at origin. Our emphasis is on the AUC Type-3 (yellow plane) as it is the only score that tackles the center-bias in eye movement data (See Fig.1 main text). We also investigate the accuracy of the object-map and saliency models on the original saliency maps using all evaluation scores (pink plane). Note that similar to Fig.1 main text, object-map is significantly better than the ITTI model with small amount of map smoothing, but with further smoothing there is no significant difference between these two models (using AUC Type-3; yellow plane). There is no significant difference between object-map and ITTI* models. Some models (AIM, HouCVPR, HouNIPS, AWS, PQFT, and SEO) outperform the object-map model in all cases. ITTI98 model is significantly better than object-map model except at the first smoothing level. Using AUC Type-3, out of 77 tested cases (yellow plane), object-map is significantly better than the early saliency only at 3 cases (original maps and 1st and 2nd smoothing cases using the sparse ITTI* model).

Appendix A: Illustration of Data and Model Maps

Fig. S4 shows sample images from *Uncommon Places* dataset (Shore et al., 2004), average saccade position over all images, and the histogram of normalized object sizes. As can be seen, the majority of annotated objects occupy less than 10% of the image. Object size is important in model performance because smaller objects could tell more about the fixation location than larger ones. The accumulation of large-size annotations leads to less dense prediction maps. Please refer to the original paper by Einhäuser et al. (2008) for more details on their model, data, and experimental setup.

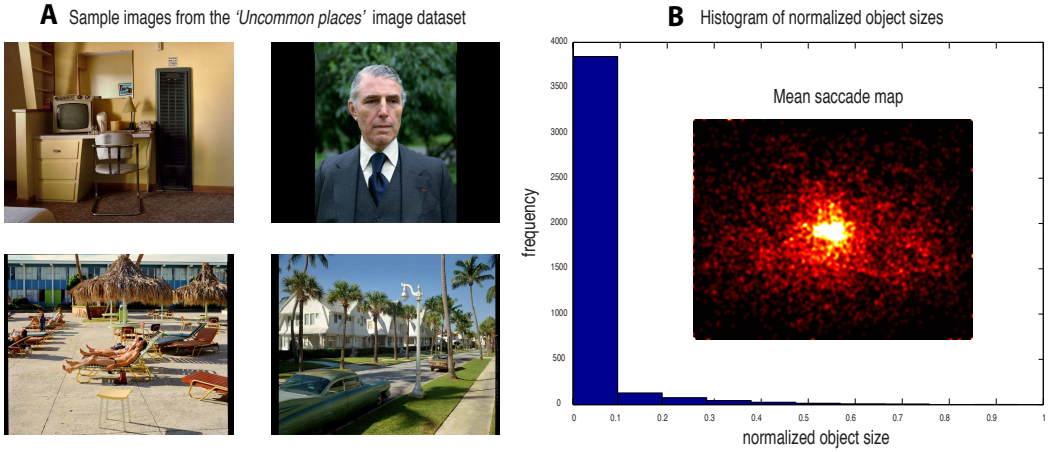

Figure S4: **A)** Sample images from the *Uncommon Places* image dataset (Shore et al., 2004). There are 93 images collected as a visual diary and come across as casual snapshots of everyday scenes. We use the same eye movement dataset used by Einhäuser et al. (2008) where you can find the experimental setup for eye movement data gathering, subjects, etc. Images have resolution of 1024 $\times$ 768 pixels and were viewed by 8 volunteers (6 male, 2 female; mean age: 23) for 3 seconds. The images were presented on a 20-inch CRT monitor, located in a dark room at 80 cm from the observer, and thus subtended 29 $\times$ 22 degrees of visual angle. A non-invasive infrared Eyelink-1000 (SR Research, Osgoode, ON, Canada) system monitored eye position at a 1000-Hz sampling rate. Thresholds to detect saccades were set to a velocity of 35°/s and an acceleration of $9500°/s^2$. The authors marked the outlines of the objects named by the observers which were verified by additional observers. **B)** Histogram of normalized annotated object size (i.e., object area divided by image area). More than 90% of bounding boxes have sizes smaller than 10% of the image. Inset: Mean saccade position over all 93 images by all subjects. There is a central peak at the image center indicating center-bias in data.

Appendix B: Computational Saliency Models

The bottom-up approach of Itti *et al.* is based on contrasts of intrinsic image features such as color, orientation, intensity, flicker, motion and others. Later implementations of this model have added channels for newer features, including text and face (Cerf et al., 2007, 2009; Judd et al., 2009), symmetry (Kootstra & Schomaker, 2009), gist and horizontal lines (Oliva & Torralba, 2001; Torralba et al., 2006), and optical flow (Fukuchi et al., 2010). These additional channels have been able to account for an increasing fraction of eye fixations. Further, several other bottom-up models with significantly different inspirations – either biologically-inspired or purely computational – have been proposed, including: Bayesian models (e.g., Surprise (Itti & Baldi, 2006), SUN (Zhang et al., 2008)), Discriminant saliency models (Gao et al., 2009), Information-theoretic models (e.g., AIM (Bruce & Tsotsos, 2009)), rarity models (Mancas, 2007), incremental coding length models by (Hou & Zhang, 2008), Spectral analysis models (e.g., PQFT (Guo & Zhang, 2010), Adaptive Whitening Saliency (AWS) (Garcia-Diaz et al., 2012a,b), and the model by Hou & Zhang (2007)), Bottom-up graphical models (e.g., GBVS (Harel et al., 2006); E-Saliency (Avraham & Lindenbaum, 2010), and the model by (Pang et al., 2008)), and Classification-based approaches (e.g., Judd et al., 2009; Kienzle et al., 2009). Because each of these models, when published, typically outperforms all others on some particular dataset and task, no one model is strictly superior to the others. Thus, instead of drawing conclusions based on only one model, we have applied several models to data of an example study. We select those models that: 1) have scored well on previous comparisons between model and human gaze allocation in free-viewing tasks, 2) are based on purely bottom-up features (not those using conceptual features such as objects or faces such as Judd et al., 2009). These models are based on well-established but different theories of human attention and visual representation. Eventually, we choose 10 models that fulfill the above two criteria: AIM, AWS, GBVS, HouCVPR and HouNIPS, ITTI98 and ITTI, PQFT, SEO, and SUN. The ITTI model (Itti & Koch, 2000) is similar to the ITTI98 (Itti et al., 1998) model but uses an iterative half-rectifying normalization operator which yields very sparse saliency maps (See also Itti & Koch, 1999).

Fig. S5.A shows a sample prediction map for each model. The average prediction map over all 93 images for each saliency model (Appendix B) is shown in Fig. S5.B. This map shows a center-bias for object-map models indicating that objects often occur at the image center. For saliency models, it seems that the average map is slightly lower at the horizontal line.

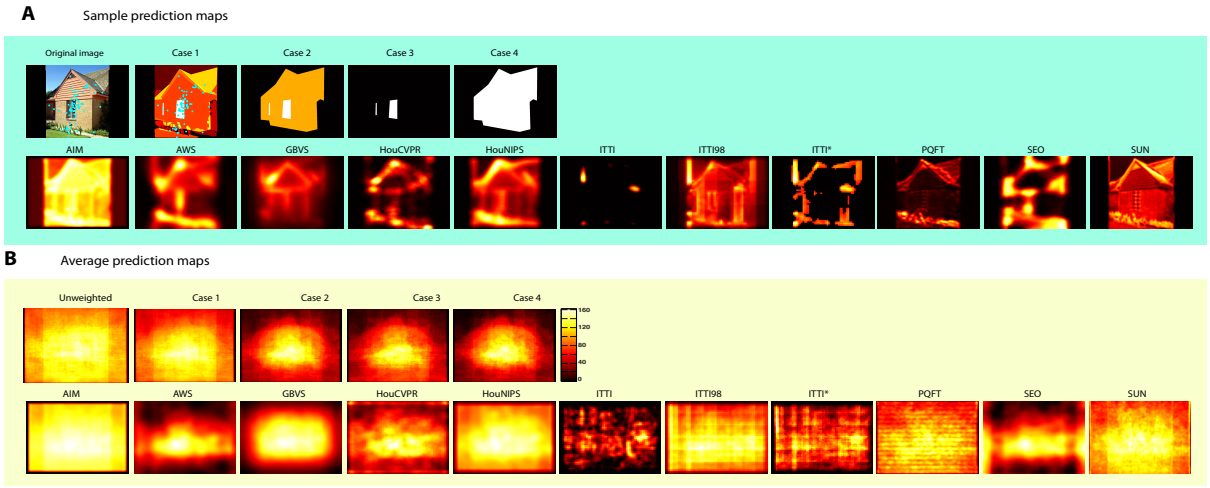

Figure S5: **A)** A sample image with eye fixations overlaid. Sample prediction maps of Einhäuser et al. (2008) model (case 1) and our implementations (cases 2 to 4), as well as 10 major bottom-up saliency models, **B)** Average prediction maps of models. The unweighted map is just the simple addition of object annotations.

Appendix C: Evaluation Scores

Here we focus on three popular scores used by many studies in the past. These scores are easy to understand and have interesting properties like well-defined bounds (lower-, upper-, and chance-level).

Area Under Curve (AUC). The most widely used score for saliency model evaluation is the Area Under the ROC Curve (Green & Swets, 1966; Tatler, 2007). Having its roots in signal detection theory, AUC measures the ability of a saliency map in separating fixated locations from non-fixated locations. Thus far three different variations of AUC exist in the literature:

AUC Type 1. First, the prediction map is resized to the image size where fixations have been recorded. Then, human fixations are considered as the positive set and some points from the image are randomly sampled (using a 2D uniform distribution) as the negative set. To form the negative set, some researchers (Itti, 2005; Berg et al., 2009) use the non-fixated locations while some take random sample from the entire image Einhäuser et al. (2008). The saliency map S is then treated as a binary classifier to separate the positive samples from the negatives. By thresholding over the saliency map, the *true positive rate* is the proportion of fixations above a threshold while the *false positive rate* is the proportion of random points above that same threshold. Then, ROC is plotted by sweeping a threshold from 0 to 1 (on a normalized map) and then AUC is calculated. Perfect prediction corresponds to a score of 1 while a score of 0.5 indicates chance level. This definition has been used in Bruce & Tsotsos (2009); Cerf et al. (2009); Berg et al. (2009); Einhäuser et al. (2008).

AUC Type 2. Here, instead of using random points, *true positive rate* (number of fixations falling on the thresholded saliency map) is plotted against the normalized saliency map area above a threshold. This score has been used by Judd et al. (2009); Ehinger et al. (2009). To tackle center-bias, Ehinger et al. (2009) proposed a control strategy (called "cross-image control"): For each saliency map, instead of using fixations for that image, they used fixations from another randomly selected image (to see how much just viewing strategy scores).

AUC Type 3. Above AUC definitions receive many true positives for a trivial central Gaussian model since majority of fixations happen in the center (Tatler, 2007). To tackle the center-bias issue and handle the spatial priors in viewing (compensation), Parkhurst & Niebur (2003) and Tatler (2007) suggested to draw random locations from the distribution of eye fixations. Here, we use the AUC Type-1, but instead of uniform random points for an image, we draw negative points from fixations of other observers over other images. This way, central fixations receive less credit compared with off-center (non-trivial) saccades. This score is also known as the Shuffled AUC score and has been used in Zhang et al. (2008); Tatler (2007); Hou et al. (2012). Tatler (2007) argue that it is better to draw the negative samples from the fixations of the same observer on different images to account for any *individual* biases.

AUC Type 4. This type of AUC is basically the same as type 3 but instead of drawing random points from fixations of other subjects over all other images, random points are same fixation locations as the current image but saliency values are taken from other images. While AUC type three gives score about 0.5, this type of AUC leads to score of exactly 0.5.

Normalized Scanpath Salience (NSS). First the saliency map is normalized to have zero mean and unit standard deviation. Next, the normalized salience values are extracted from each point corresponding to the fixation locations along a subject's scanpath, and the mean of these values, or NSS (Peters et al., 2005; Parkhurst et al., 2002), is taken as a measure of the correspondence between the saliency map and the scanpath. Mathematically, NSS is:

$$NSS = \frac{1}{N} \sum_{i=1}^{N} \frac{S(x_g^i, y_g^i) - \mu_S}{\sigma_S} \tag{2}$$

where $S(x_g^i, y_g^i)$ is the saliency value at the i-th human eye position (x_g^i, y_g^i), N is the number of fixations for each image, and μ_S and σ_S are mean and standard deviation of the saliency map, respectively. $NSS = 1$ indicates that the subjects' eye positions fall in a region whose predicted density is one standard deviation above average. Meanwhile, $NSS \leq 0$ indicates that the model performs no better than picking a random position.

Linear Correlation Coefficient (CC). The third score is the correlation coefficient between human saliency map G (a map with frequency of saccades at each location which is usually convolved with a small Gaussian kernel; here of size $0.28^\circ \times 0.28^\circ$; (See Fig.1 main text) and a model's saliency map S (Jost et al., 2005; Rajashekar et al., 2008):

$$CC(G,S) = \frac{\sum_{xy}(G(x,y) - \mu_G).(S(x,y) - \mu_S)}{\sqrt{\sigma_G^2.\sigma_S^2}} \tag{3}$$

In above formula, μ and σ^2 are the mean and the variance of the values in these maps and $\sum_{xy}$ is the covariance matrix. An interesting advantage of CC is its capacity to compare two variables by providing a single scalar value between -1 and +1. Another advantage of CC is that it has the well-defined upper-bound of 1. When CC is close to $+1/-1$ there is almost a perfect linear relationship between the two variables.

Note that CC, NSS, and classic AUC scores (Types 1 and 2) are all affected by center-bias. Here, we emphasize more on the Shuffled AUC score which is becoming a standard score for saliency model evaluation (Zhang et al., 2008; Borji et al., 2012; Hou et al., 2012).

Results using CC and NSS are shown in Fig. S6. Using both scores and original maps, the object-map model is significantly above the ITTI and ITTI* while being significantly below other models (paired t-test, $\alpha < 0.0045$; Bonferroni-corrected). This is because generated maps by the ITTI model are very sparse. GBVS model takes advantage of center-bias in data implicitly which results in high CC and NSS scores for this model. Models with more blurry/smoothed maps achieve higher CC values. MEP and Normal random model outperform others consistent with the AUC Type-1. Since these two models have peaks at locations that many people look at (usually center), they achieve very high NSS scores. T-test p-values (Bonferroni corrected) of object-map model versus ITTI, ITTI98, and AWS model in order are: $p = 2.33 \times 10^{-7}, 0.517$, and 1.22×10^{-7}. These values for NSS are: $p = 1.5 \times 10^{-6}, 0.4646$, and 4.70×10^{-8}. Based on current results and our previous investigations (Borji et al., 2012), the AWS model has the highest fixation prediction power among saliency models (averaged over all different scores). Therefore, it is thus far the best predictor of human fixations and we suggest its use in future studies of visual attention. Of course Einhäuser et al. were not aware of this model at the time of their study (which is why, in the paper, we continued to use the ITTI and ITTI98 models that were available at the time). Similar to the ITTI98 model, AWS uses basic features (luminance, and a, b color channels from the *Lab* image representation) in several scales. It then decorrelates the multi-scale responses and extracts a local measure of variability for saliency calculation. A more conservative approach will be using multiple high-performing models (i.e., GBVS, AIM, and HouNIPS).

By adding center-bias (Fig. S6.B), usually CC and NSS scores rise to a maximum value and then drop (see also Borji et al., 2012). Using CC, the object-map model is significantly below the ITTI model with the first ($p = 0.00017$) and second Gaussian kernels ($p = 0.00026$). There is no significant difference between these two models after adding the third Gaussian kernel to them (paired t-test, $\alpha < 0.05/5 = 0.01$). There is also no significant difference between object-map and ITTI98 models using first and third Gaussian kernels. The AWS model outperforms the object-map model in all cases with $p = 1.3402 \times 10^{-8}, p = 6.6647 \times 10^{-11}$, and 3.4256×10^{-11} for the three Gaussian kernels, respectively. Nearly the same pattern holds for the NSS score where the object-map model is significantly lower than the ITTI model with the first and second Gaussian kernels, but not the third kernel, for which there is no significant difference. Again, there is no significant difference between ITTI98 and object-map models. Using CC and NSS scores, AWS model outperforms the object-map model significantly with a large margin. The AWS model is significantly above the object-map model using all three Gaussian kernels. The object-map model is significantly above the ITTI* model only using the first Gaussian kernel.

With smoothing (Fig. S6.C), the object-map model is significantly above the ITTI model using CC and NSS scores and using all three Gaussian kernels (for smoothing). There is no significant difference between ITTI98 and the object-map models using all three Gaussian kernels. AWS model outperforms the object-map model significantly using all three Gaussians. Similar to adding center-bias, the object-map model is significantly above the ITTI* model only using the first Gaussian kernel.

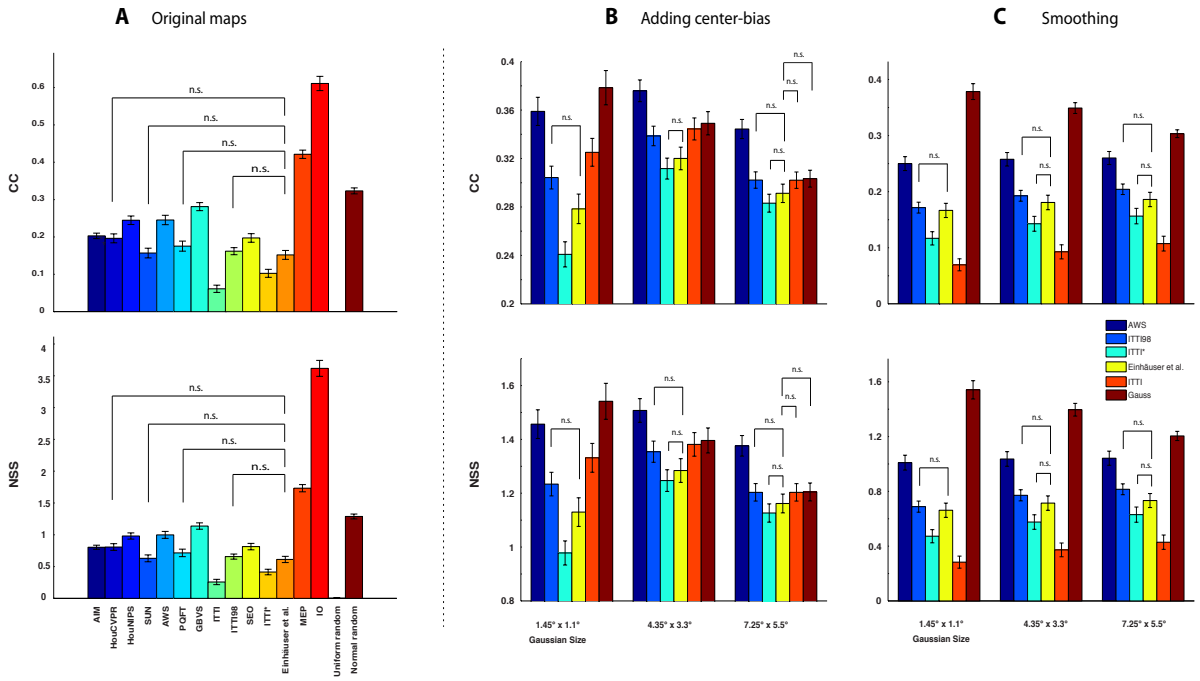

Figure S6: **A)** Correlation coefficient (CC) and Normalized Scanpath Saliency (NSS) scores of models. Again aligned with our conclusions from Fig. S2, the object-map model is significantly above the ITTI but not ITTI98 model. Object-map is however significantly below several newer saliency models using these two scores (e.g., AIM, HouCVPR, HouNIPS, AWS, and GBVS). The Inter-observer model stands on top of all models. Note that these two scores result in high values for MEP leading it above other models with a large margin. GBVS that uses center-bias is getting high scores here. Significance values are Bonferroni-corrected (paired t-test, $\alpha < 0.0045$). **B and C)** Prediction accuracies of models using CC and NSS scores: B) center-bias added and C) smoothed saliency maps. With adding center-bias using CC and NSS, object-map model does not outperform any model. Smoothing raises accuracy of models using CC and NSS scores (except the Gauss model). There is no significant difference between ITTI* and the object-map with adding center-bias and smoothing. Significance values are according to Bonferroni-corrected paired t-test ($\alpha < 0.05/5 = 0.01$).

Appendix D: The Object-based Models

We devised three other variants of Einhäuser et al. (2008)'s model, in addition to their original proposal, to cover all bases and to gauge the dependency of conclusions on the exact model (similar to what we do with saliency models). These cases are motivated and described below:

Case 1) By analyzing Einhäuser *et al.*'s data, we noticed that their subjects often tended to report the same object as the first remembered one. To quantify subject agreement on the $m-th$ image (i.e., a^m), we define the following variable r^m_{uv} which is one when both $u-th$ and $v-th$ subjects remembered the same object as the first one on image m, and 0 otherwise (i.e., $r^m_{uv} = (o^m_u == o^m_v)$ where o^m_u and o^m_v are the first remembered objects of subjects u and v over image m). Then, we loop through all pairs of subjects and calculate the histogram of these values:

$$a^m = \frac{2}{K(K-1)} \sum_{u=1}^{K-1} \sum_{v=u+1}^{K} r^m_{uv} \tag{4}$$

The above score has the well-defined lower-bound of 0 when there is no agreement between subjects and upper-bound of 1 when all subjects remembered the same object at the first place. Fig. S7.A suggests that the histogram of a values has a rightward shift compared with a process which randomly selects a value for each of the o^m_u and o^m_v variables from the set of annotated objects for image m.

Having seen this, we build a map out of those first recalled objects. Formulation is as in case 1, but w_{ij} is 1 if object j is the first remembered object by subject i and 0 otherwise.

Case 2) Another possibility is that subjects may look at the frequently remembered objects. Here, we measure the agreement of subjects by the most frequently remembered object. For image m, we calculate $b_m = t_m/8$ which is the fraction of subjects that recalled the most remembered object on this image (t_m). A value close to 1 means high recall consistency. The histogram of b values over all images is shown in Fig. S7.B. As it shows human agreement is higher than random. We simulate a random process as follows: generate a random binary table where each element indicates whether an object was remembered by a subject or not (rows are subjects and columns are the annotated objects for an image). Find the object occurrence in each column, divide it to K (number of subjects), and then pick the maximum.

High recall consistency among subjects leads us to propose the third variant. We build an object map containing only the annotation of the object that has been recalled the most: $object-map = B_e,\ e = \arg\max_j\ n_j$ where n_j represents the frequency in which object j has been recalled over all subjects.

Case 3) Similar to the case 3, here we calculate the histogram of the subject agreement for the most likely first remembered object. Again, according to Fig. S7.C, humans agree with each other more often than chance. The random process is similar to the third case with the exception that there is only one non-zero element in each row of the table (i.e., randomly assigning a first remembered object). Thus, we propose the forth variant. Here, the prediction map is similar to case 3, with the only difference that e is the index of the object that has been maximally remembered first (i.e., $e = \arg\max_j\ n_j$ where n_j is the number of times object j is remembered first).

We test whether other cases of the object-map model could predict fixations better than early saliency. Results are shown in Fig. S7 using AUC Types 1 and 3. We find that case-1 (the map weighted by the object recall frequency) performs significantly above other cases (paired t-test; $\alpha < 0.0125$). Thus, since case-1 was not performing better overall than many saliency models in Fig. S2, we conclude that object-based models (weighted recall of objects, first remembered, max first remembered, or max remembered object) cannot account for eye fixations better than the early saliency although they all perform significantly above chance.

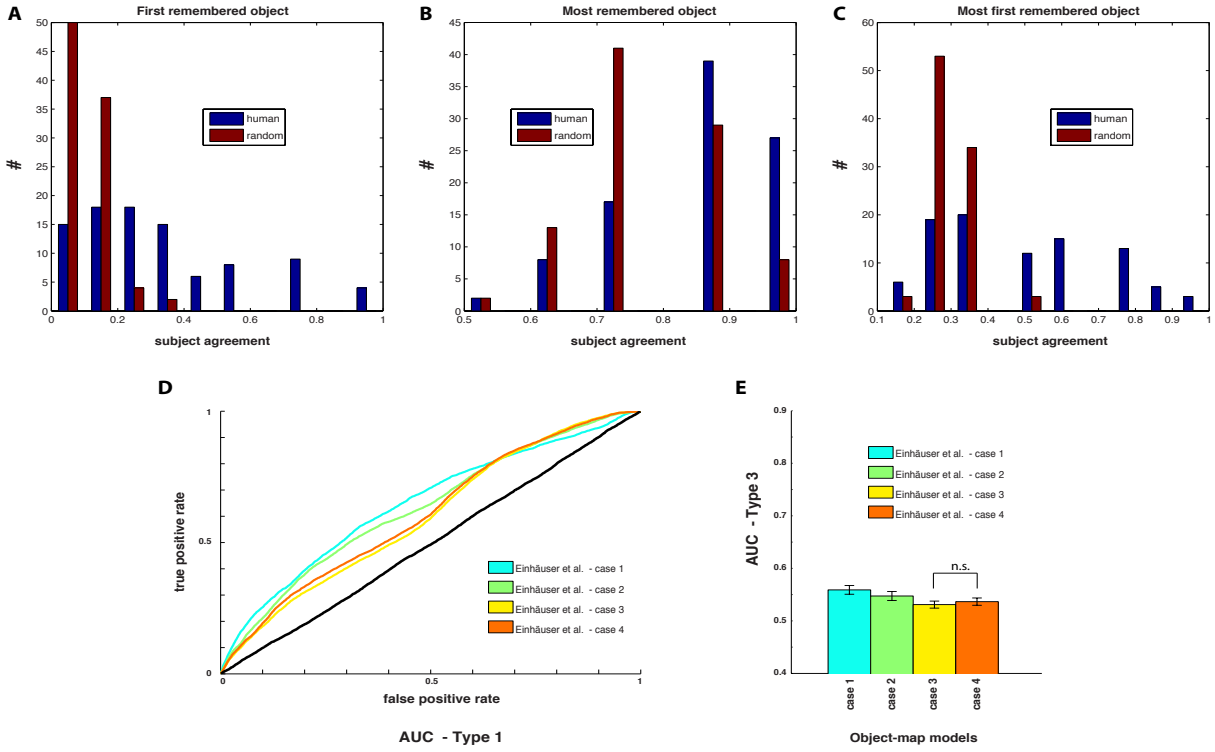

Figure S7: Histogram of subject agreements for *the first remembered object* (A), *the most remembered object* (B), and the *most first remembered object* (C). Blue bars represent agreements of humans and red bars are for the random processes. In all three histograms, human agreements are higher than chance. Comparison of accuracies of four cases of the object-based models over Einhäuser et al. (2008) data (D and E). As it shows, case 1 is the best among all models (significantly above others). Object-map case 3 scores the lowest among four cases. All of the 4 cases are significantly above the ITTI model but significantly below the ITTI98. The fact that all cases are above chance indicates that objects convey some information regarding eye fixation positions. Every difference is statistically significant except between cases 3 and 4.