What do deep saliency models learn about where we look in scenes?

Taylor R. Hayes (trhayes@ucdavis.edu)
University of California, Davis
John M. Henderson
University of California, Davis

Research Article

Keywords: deep saliency models, cognitive theories of attention, image features, scene meaning.

Posted Date: April 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-420396/v1

What do deep saliency models learn about where we look in scenes?

Taylor R. Hayes[1*] **and John M. Henderson**[1,2]

[1]Center for Mind and Brain, University of California, Davis, 95618, USA
[2]Department of Psychology, University of California, Davis, 95616, USA
[*]trhayes@ucdavis.edu

ABSTRACT

Deep saliency models represent the current state-of-the-art for predicting where humans look in real-world scenes. However, for deep saliency models to inform cognitive theories of attention, we need to know *how* deep saliency models predict where people look. Here we open the black box of deep saliency models using an approach that models the association between the output of 3 prominent deep saliency models (MSI-Net, DeepGaze II, and SAM-ResNet) and low-, mid-, and high-level scene features. Specifically, we measured the association between each deep saliency model and low-level image saliency, mid-level contour symmetry and junctions, and high-level meaning by applying a mixed effects modeling approach to a large eye movement dataset. We found that despite different architectures, training regimens, and loss functions, all three deep saliency models were most strongly associated with high-level meaning. These findings suggest that deep saliency models are primarily learning image features associated with scene meaning.

Introduction

Our everyday visual world contains too much information to be taken in at once so we filter our visual world by moving our eyes to prioritize some regions over others. But how do humans know where to look to efficiently build a representation and understanding of complex, real-world scenes? One approach to answering this question is to construct computational models that predict where people look in scenes. Deep convolutional neural network models of saliency (i.e., 'deep saliency models') reflect the current state-of-the-art computational models for predicting human fixations in scenes[1]. Although these models often generate very good predictions of human behavior, relatively little is known about *how* they predict where people look. To use deep saliency models to inform cognitive theories of attention, we will need a better understanding of what deep saliency models are learning about where people look in scenes.

To begin, it is helpful to distinguish *deep saliency models* from *image saliency models*. Image saliency models are computed from the scene image alone by combining local contrasts in low-level, pre-semantic image features like color, luminance, and orientation across multiple spatial scales[2–6]. For example, a bright red flower surrounded by green grass would be a region that would be predicted by an image saliency model to capture attention. In comparison, deep saliency models use a data-driven approach that combines deep convolutional neural networks trained on large object recognition datasets (e.g., VGG-16 or VGG-19[7]) with additional network layers that are subsequently trained on human fixation data[8]. Within this approach, deep saliency models learn a mapping between the pre-trained object recognition features and the human fixation data they are trained on. Therefore, a critical difference between these two classes of saliency models is that image saliency models only use low-level image features to generate their predictions, whereas deep saliency models likely use some combination of low-, mid-, and high-level features to generate their predictions because they are trained using both object recognition and human fixation data[9]. Therefore, in order to understand the factors that drive deep saliency models, we will need to assess the association between low-, mid-, and high-level scene information and deep saliency model performance.

A large body of previous research has shown an association between low-level stimulus features and attention. Early theories of attention focused on the role of low-level feature differences in capturing attention and were based on experiments using simple stimuli like lines and/or basic shapes that varied in low-level features like orientation, color, luminance, texture, shape, or motion[10–12]. These early theories were subsequently formalized into computational image 'saliency' models that pooled contrasts in low-level feature maps (e.g., luminance, color, orientation) based on mechanisms observed in early visual cortex such as center-surround dynamics to generate quantitative predictions in the form of 'saliency maps'[4,5,13–15]. Image saliency maps were shown to be significantly correlated with where people looked in scenes[2–6]. This foundational work spawned a large number of image saliency models (e.g., Graph-based saliency model[3]; Adaptive Whitening Saliency[16]; RARE[17], Attention based on Information Maximization[18]) that each generate image saliency maps in different ways to improve their overall biological plausibility and/or performance on benchmark datasets[1]. Given the extensive theoretical, biological, and

computational work on the role of low-level features in guiding attention, it will be important to quantify the degree to which low-level features are associated with deep saliency model performance.

Mid-level vision is thought to play a role in organizing low-level features in specific ways (e.g., Gestalt principles) that facilitate higher-level recognition processes[19–23]. However, there has been very little work on the role that mid-level features play in guiding overt attention in scenes[9]. A recent study[9], showed that two different proposed mid-level features, local symmetry and contour junctions, contributed to category-specific scene attention during a scene memorization task in grayscale scenes and line drawings. The mid-level scene category predictions were also computed over discrete temporal time bins, and the results suggested that symmetry contributed to both early bottom-up and later top-down guidance, while junctions contributed mostly to later top-down guidance[9]. Therefore, in the present work, it will also be useful to quantify the association between mid-level features (i.e., local symmetry and contour junctions) and deep saliency models.

Finally, there is a growing literature suggesting that high-level semantic density plays an important role in guiding attention in real-world scenes[24–30]. Much of this work has shown that high-level semantics often overrides low-level salience[25–28,31]. While many semantic scene studies manipulate a single or small number of objects in each scene, it is also possible to use human raters to rate the meaningfulness of scene regions to generate a continuous distribution of semantic density across the entire scene (i.e., meaning map)[26]. Meaning maps have been shown to be one of the strongest predictors of where people look in scenes across a wide variety of viewing tasks including scene memorization[26,27], visual search[32], free viewing[33], scene description[34], and saliency search[35]. Therefore, in the present study it will be important to assess the degree to which the image features learned by deep saliency models are associated with high-level meaning.

In the present work, we had two main goals. First, we sought to replicate and extend recent results demonstrating that prominent deep saliency models (MSI-Net[36]; DeepGaze II[37]; and SAM-ResNet[38]) provide excellent predictions of human attention during free-viewing of scenes. We addressed this goal using a large eye movement dataset based on 100 scenes and 100 where participants performed active scene viewing tasks rather than free-viewing. Our analyses explicitly accounted for center bias[39,40] and the random effects of viewer and scene using a mixed effects modeling approach[30,41]. Second, and more importantly, we investigated the nature of the scene features that deep saliency models use to predict attention by modeling the association between attention, deep saliency model output, and low-level saliency[3,42], mid-level symmetry and junctions[43,44], and high-level scene meaning[26].

Results

We first examined how well each deep saliency model predicted where subjects did and did not look in the scenes they viewed. Specifically, we fit a logistic GLME model for each deep saliency model in which whether a region was fixated or not (Fig. 1a) was modeled as a function of the fixed effect of a scene region's mean deep saliency model value (Fig. 1b, c, d), center proximity value (Fig. 1e), and their interaction. Subject and scene were treated as random intercepts. As shown in Fig. 2a, each deep saliency model had a significant positive association with where observers looked (MSI-Net $\beta = 2.19$, CI $[2.17, 2.20]$, $z = 295.4$, $p < .001$; DeepGaze II, $\beta = 1.82$, CI $[1.81, 1.83]$, $z = 339.9$, $p < .001$; SAM-ResNet, $\beta = 2.57$, CI $[2.55, 2.59]$, $z = 242.8$, $p < .001$). Additionally, each model interacted with center proximity (MSI-Net, $\beta = -0.15$, CI $[-0.16, -0.13]$, $p < .001$; DeepGaze II, $\beta = 0.16$, CI $[0.15, 0.17]$, $p < .001$; SAM-ResNet, $\beta = -0.22$, CI $[-0.24, -0.20]$, $p < .001$) visualized as a function of fixation probability in Fig. 2d. All three models did comparable jobs of predicting whether a scene region would be fixated or not (MSI-Net=0.82, DeepGaze II=0.83, , SAM-ResNet=0.81). Taken together our results replicate previous findings[36–38] and establish that MSI-Net, DeepGaze II, and SAM-ResNet also predict scene attention well in active viewing tasks.

However, showing that deep saliency models are strongly associated with where people look in scenes during active viewing, does not tell us anything about *how* these models predict where people will look in each scene. Therefore, to gain insight into how the deep saliency models are predicting, we turned the analysis on its head and modeled the associations between the deep saliency model values and the corresponding low-, mid-, and high-level map values for each fixated scene region. Specifically, we fit a linear mixed effects model for each deep saliency model (Fig. 1b, c, d), where we modeled the fixated deep saliency model values as a function of the low-level (image saliency, Fig. 1f), mid-level (symmetry, Fig. 1g and junctions, Fig. 1h), high-level (meaning, Fig. 1i), and center proximity maps (Fig. 1e). Subject and scene were again treated as random intercepts in each model. This approach allows us to measure the association of each deep saliency model (Fig. 1b, c, d) with each of our defined feature maps (Fig. 1e, f, g, h, i) while controlling for the random effects of observers and scenes, and addresses our main question of interest, what do deep saliency models learn about where we look in scenes?

The feature association results are shown below for MSI-Net (Fig. 3 and Table 1), DeepGaze II (Fig. 4 and Table 2) and SAM-ResNet (Fig. 5 and Table 3. Looking across all three models, the most consistent finding was that high-level meaning was most strongly associated with the deep saliency model output (MSI-Net $\beta = 0.33$, CI $[0.33, 0.34]$, $t = 153.1$, $p < .001$; DeepGaze II, $\beta = 0.44$, CI $[0.44, 0.45]$, $t = 288.5$, $p < .001$; SAM-ResNet, $\beta = 0.28$, CI $[0.27, 0.29]$, $t = 93.2$, $p < .001$). Second, we also consistently observed that low-level saliency was associated with each deep saliency model

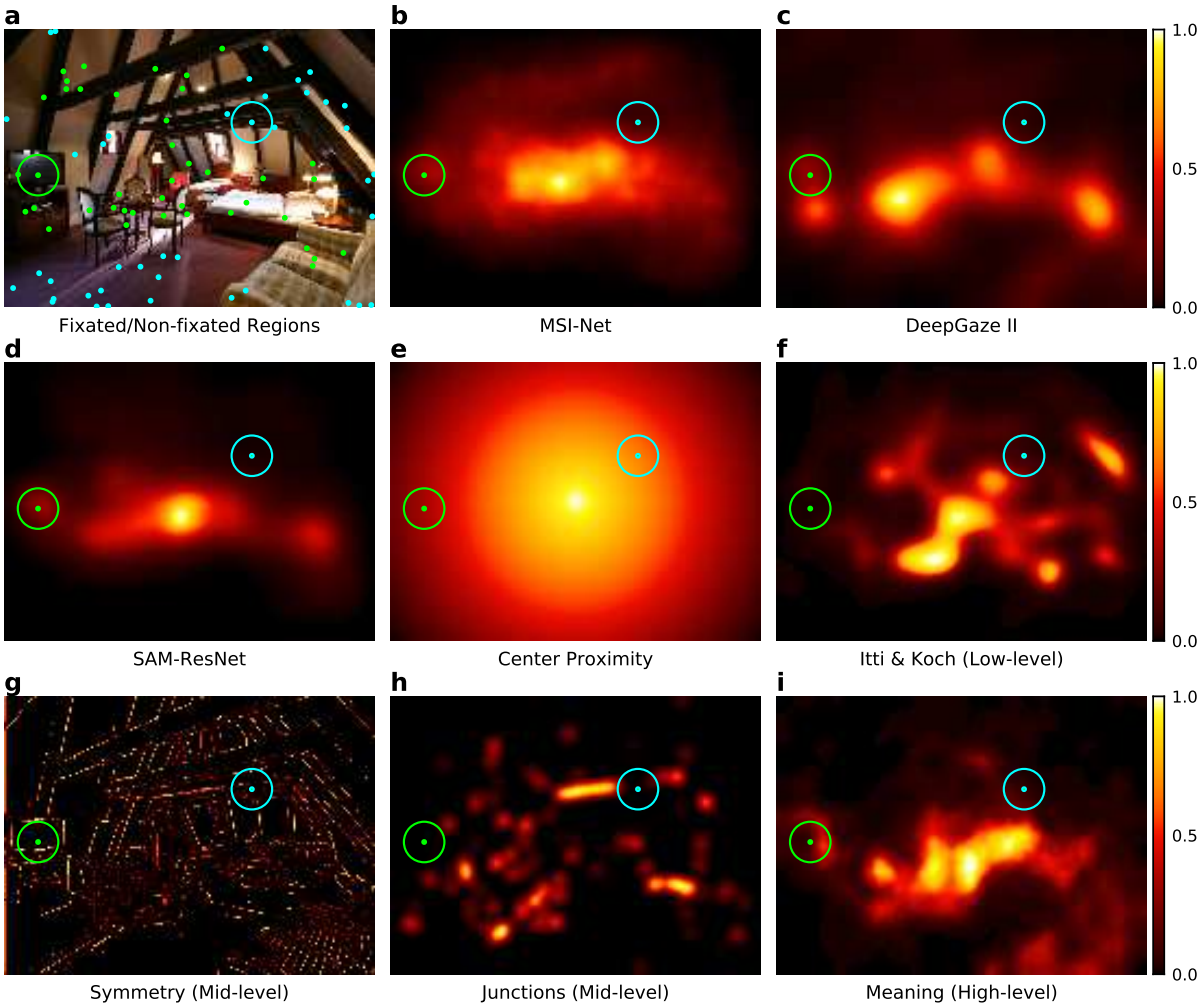

Figure 1. *Scene with the fixated and non-fixated regions for a single subject, and the corresponding deep saliency and feature maps.* **a** The green dots show the fixation locations for a single viewer and the cyan dots indicate randomly sampled non-fixated regions that represent where this subject did not look in this scene. Together these locations provide an account of the regions in this scene that did and did not capture this subject's attention. Each fixated and non-fixated location was then used to compute a mean model value for each deep saliency model map (**b, c, d**) and feature map (**e, f, g, h, i**) across a 3° window (shown as circles around one example fixated and non-fixated location).

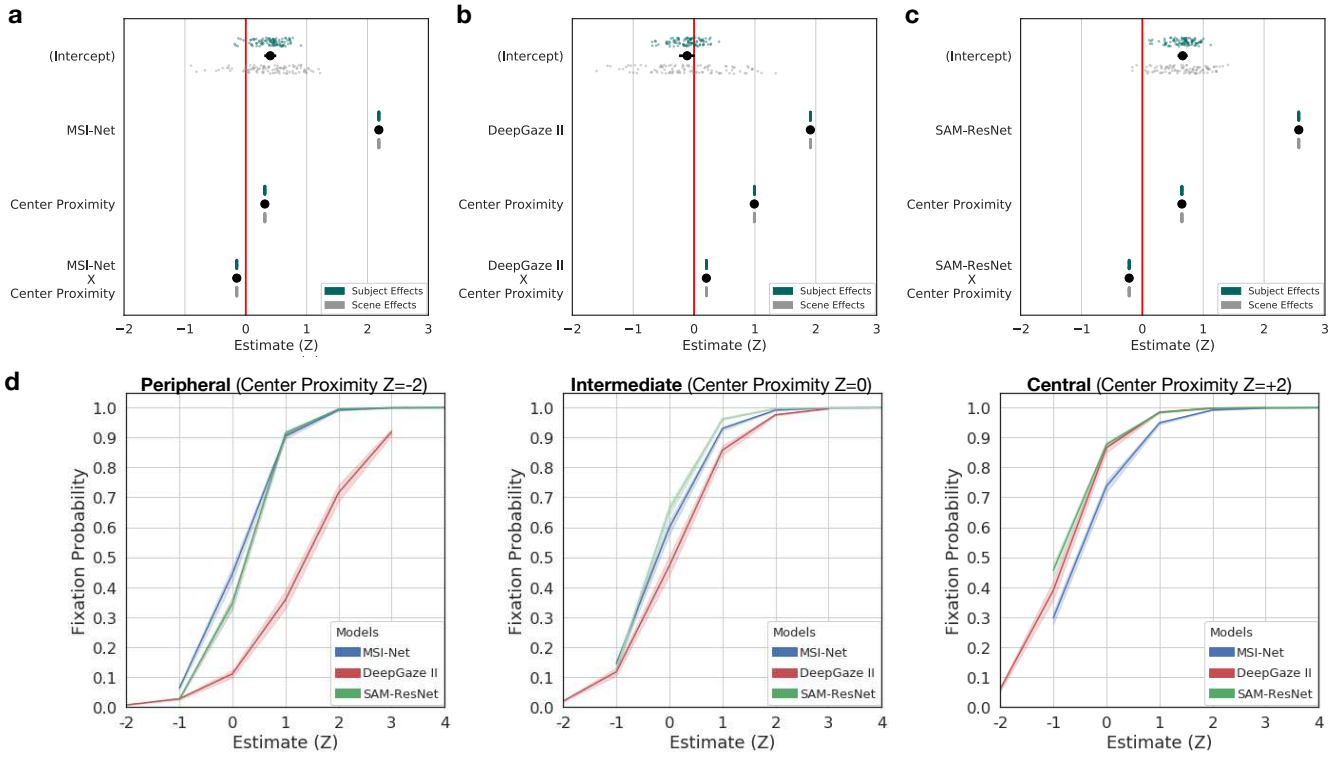

Figure 2. *Deep saliency model general linear mixed effects model results.* Whether a scene region was fixated or not was modeled as a function of the deep saliency model value, center proximity value, and their interaction as fixed effects (**a** MSI-Net; **b** DeepGaze II; **c** SAM-ResNet). The black dots with lines show the fixed effect estimates and their 95% confidence intervals. Subject (green dots) and scene (grey dots) were both accounted for in the model as random intercepts. **d** A line plot of the interaction between center proximity (panels) and each deep saliency model (colored lines) as a function of fixation probability. All error bands reflect 95% confidence intervals.

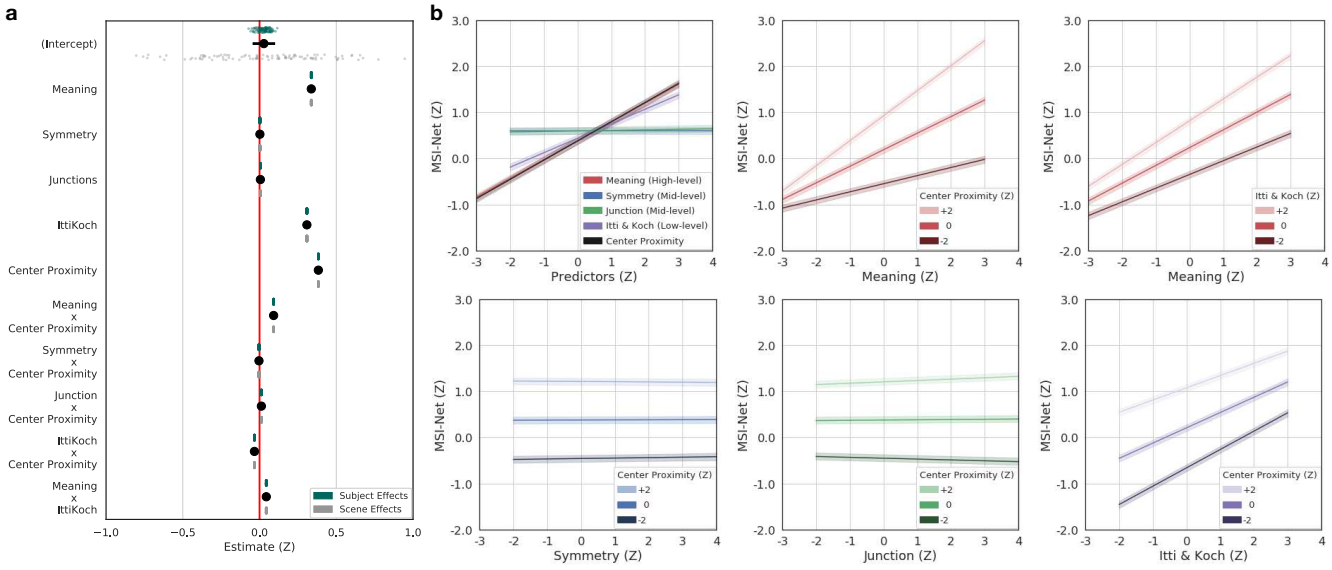

Figure 3. *MSI-Net linear mixed effects model, marginal effects, and interactions.* **a** Fixated MSI-Net values as a function of low-, mid-, and high-level features and interactions. The blackdots with lines show the fixed effect estimates and their 95% confidence intervals. Subject (greendots) and scene (grey dots) were both accounted for in the model as random intercepts. **b** Line plots of all model marginal effects and all model interactions. All error bands reflect 95% confidence intervals.

	Fixed effects					Random effects, *SD*	
Predictors	β	95% CI	*SE*	*t*-value	*p*	by-subject	by-scene
Intercept	0.03	[-0.04 0.10]	0.04	0.75	0.45	0.04	0.37
Meaning	0.33	[0.33 0.34]	0.002	153.13	< 0.001***	-	-
Symmetry	0.002	[-0.003 0.008]	0.003	0.89	0.37	-	-
Junctions	0.006	[0.003 0.008]	0.001	4.27	< 0.001***	-	-
IttiKoch	0.308	[0.305 0.312]	0.002	166.54	< 0.001***	-	-
Center Proximity	0.383	[0.380 0.386]	0.002	241.59	< 0.001***	-	-
Meaning:Center Proximity	0.092	[0.088 0.095]	0.002	52.82	< 0.001***	-	-
Symmetry:Center Proximity	-0.004	[-0.006 -0.001]	0.001	-2.81	< 0.01**	-	-
Junction:Center Proximity	0.012	[0.010 0.014]	0.001	10.28	< 0.001***	-	-
IttiKoch:Center Proximity	-0.034	[-0.036 -0.030]	0.002	-21.64	< 0.001***	-	-
Meaning:IttiKoch	0.044	[0.041 0.047]	0.002	25.86	< 0.001***	-	-

Table 1. *MSI-Net LME Results.* Beta estimates (β), 95% confidence intervals (CI), standard errors (SE), $t-$values, and *p*-values (*p*) for each fixed effect and standard deviations (*SD*) for the random effects of subject and scene.

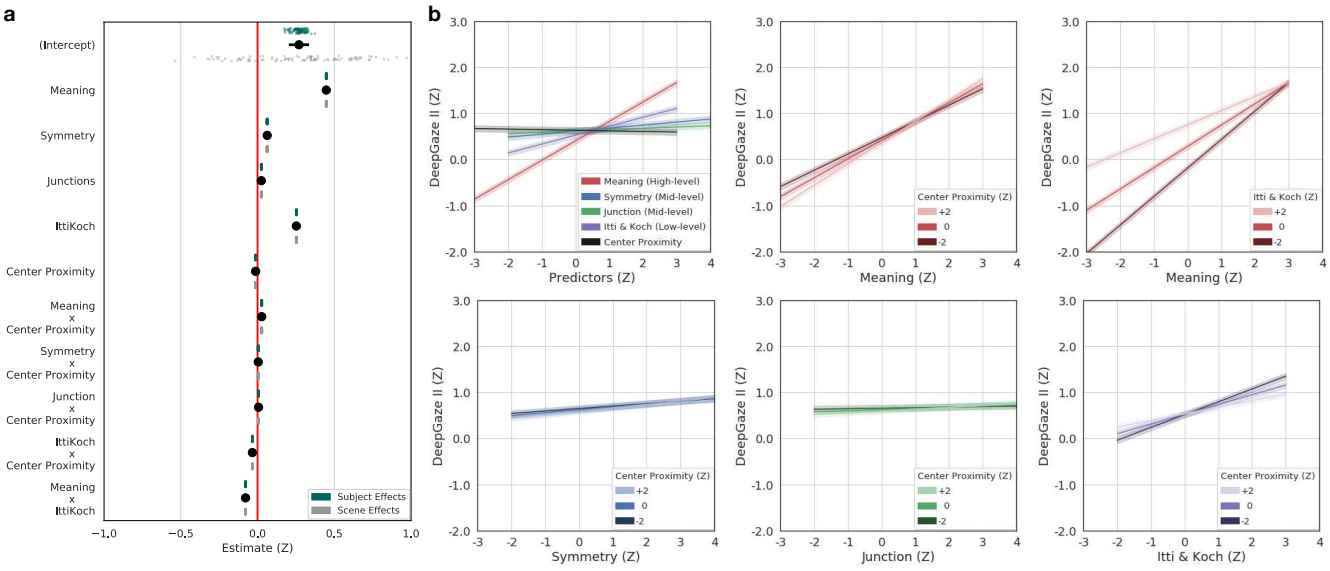

Figure 4. *DeepGaze II LME model, marginal effects, and interactions.* **a** Fixated DeepGaze II values as a function of low-, mid-, and high-level features and interactions. The blackdots with lines show the fixed effect estimates and their 95% confidence intervals. Subject (greendots) and scene (grey dots) were both accounted for in the model as random intercepts. **b** Line plots of all model marginal effects and all model interactions. All error bands reflect 95% confidence intervals.

	Fixed effects					Random effects, *SD*	
Predictors	β	95% CI	*SE*	*t*-value	*p*	by-subject	by-scene
Intercept	0.27	[0.20 0.34]	0.03	8.00	0.13	0.04	0.33
Meaning	0.448	[0.445 0.451]	0.002	288.48	< 0.001***	-	-
Symmetry	0.06	[0.05 0.07]	0.002	33.57	< 0.01**	-	-
Junctions	0.025	[0.023 0.026]	0.001	26.83	< 0.001***	-	-
IttiKoch	0.253	[0.250 0.255]	0.001	193.45	< 0.001***	-	-
Center Proximity	-0.012	[-0.015 -0.010]	0.001	-11.16	< 0.001***	-	-
Meaning:Center Proximity	0.027	[0.025 0.030]	0.001	22.16	< 0.001***	-	-
Symmetry:Center Proximity	0.005	[0.003 0.007]	0.001	5.13	< 0.001***	-	-
Junction:Center Proximity	0.007	[0.005 0.008]	0.001	8.02	< 0.001***	-	-
IttiKoch:Center Proximity	-0.034	[-0.036 -0.031]	0.001	-30.99	< 0.001***	-	-
Meaning:IttiKoch	-0.078	[-0.081 -0.076]	0.001	-65.19	< 0.001***	-	-

Table 2. *DeepGaze II LME Results.* Beta estimates (β), 95% confidence intervals (CI), standard errors (SE), $t-$values, and *p*-values (*p*) for each fixed effect and standard deviations (*SD*) for the random effects of subject and scene.

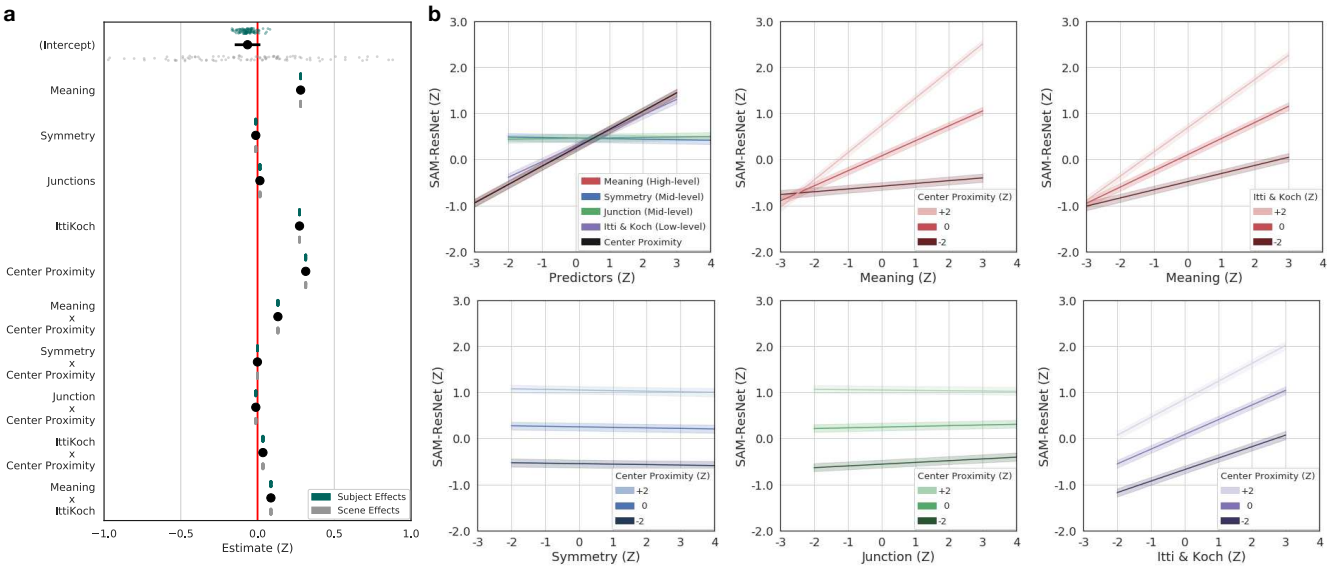

Figure 5. *SAM-ResNet LME model, marginal effects, and interactions.* **a** Fixated SAM-ResNet values as a function of low-, mid-, and high-level features and interactions. The blackdots with lines show the fixed effect estimates and their 95% confidence intervals. Subject (greendots) and scene (grey dots) were both accounted for in the model as random intercepts. **b** Line plots of all model marginal effects and all model interactions. All error bands reflect 95% confidence intervals.

	Fixed effects					Random effects, *SD*	
Predictors	β	95% CI	*SE*	*t*-value	*p*	by-subject	by-scene
Intercept	-0.06	[-0.15 0.02]	0.04	-1.53	0.13	0.05	0.42
Meaning	0.28	[0.27 0.29]	0.003	93.18	< 0.001***	-	-
Symmetry	-0.01	[-0.02 0.00]	0.004	-3.14	< 0.01**	-	-
Junctions	0.016	[0.012 0.019]	0.002	8.75	< 0.001***	-	-
IttiKoch	0.274	[0.269 0.278]	0.003	107.84	< 0.001***	-	-
Center Proximity	0.314	[0.310 0.318]	0.002	144.54	< 0.001***	-	-
Meaning:Center Proximity	0.133	[0.128 0.137]	0.002	55.82	< 0.001***	-	-
Symmetry:Center Proximity	-0.001	[-0.004 0.003]	0.002	-0.37	0.71	-	-
Junction:Center Proximity	-0.011	[-0.014 0.008]	0.002	-6.99	< 0.001***	-	-
IttiKoch:Center Proximity	0.035	[0.031 0.039]	0.002	16.72	< 0.001***	-	-
Meaning:IttiKoch	0.087	[0.083 0.092]	0.002	37.42	< 0.001***	-	-

Table 3. *SAM-ResNet LME Results*. Beta estimates (β), 95% confidence intervals (CI), standard errors (SE), $t-$values, and p-values (p) for each fixed effect and standard deviations (*SD*) for the random effects of subject and scene.

(MSI-Net $\beta = 0.31$, CI $[0.30, 0.31]$, $t = 166.5$, $p < .001$; DeepGaze II, $\beta = 0.253$, CI $[0.250, 0.255]$, $t = 193.5$, $p < .001$; SAM-ResNet, $\beta = 0.27$, CI $[0.26, 0.28]$, $t = 107.8$, $p < .001$). Finally, the mid-level junctions (MSI-Net $\beta = 0.006$, CI $[0.003, 0.008]$, $t = 4.27$, $p < .001$; DeepGaze II, $\beta = 0.03$, CI $[0.02, 0.03]$, $t = 26.83$, $p < .001$; SAM-ResNet, $\beta = 0.016$, CI $[0.012, 0.019]$, $t = 8.75$, $p < .001$) and symmetry (DeepGaze II, $\beta = 0.06$, CI $[0.05, 0.07]$, $t = 33.57$, $p < .01$; SAM-ResNet, $\beta = -0.01$, CI $[-0.02, 0.00]$, $t = -3.1$, $p < .01$) features were only weakly associated with the deep saliency model values. Together these findings suggest that deep saliency models are significantly associated with high-, mid-, and low-level saliency maps, but are most strongly associated with high-level meaning.

Based on previous work[26,27,32] that showed a relationship between high-level meaning and low-level saliency, we included an interaction term (high-level meaning X low-level image saliency) in each of our decomposed deep saliency model analyses. The high-level by low-level interaction was significant in all three deep saliency models (MSI-Net $\beta = 0.04$, CI $[0.04, 0.05]$, $t = 25.86$, $p < .001$; DeepGaze II, $\beta = -0.08$, CI $[-0.08, -0.07]$, $t = -65.2$, $p < .001$; SAM-ResNet, $\beta = 0.09$, CI $[0.08, 0.09]$, $t = 37.4$, $p < .001$), but displayed different interaction patterns. MSI-Net (Fig. 3b, top-right) and SAM-ResNet (Fig. 5b, top-right) displayed a similar interaction pattern, where as a fixated region's meaning value increased the predicted MSI-Net and SAM-ResNet values increased more quickly with greater low-level saliency. DeepGaze II displayed the opposite interaction pattern (Fig. 4b, top-right), where as a fixated region's meaning value increased the predicted DeepGaze II values increasingly were unaffected by low-level saliency. These divergent interaction patterns suggest that MSI-Net and SAM-ResNet predict a scene region is more likely to be fixated if it is both meaningful and visually salient, while DeepGaze II prediction is associated with increasingly discounting low-level saliency as a scene region becomes more meaningful.

Finally, center proximity also played a significant role in each deep saliency model both as a fixed effect and as an interaction term. The effect of center proximity was larger in MSI-Net ($\beta = 0.38$, CI $[0.38, 0.39]$, $t = 241.6$, $p < .001$) and SAM-ResNet ($\beta = 0.31$, CI $[0.31, 0.32]$, $t = 144.5$, $p < .001$) compared to DeepGaze II ($\beta = -0.01$, CI $[-0.015, -0.010]$, $t = -11.2$, $p < .001$). The interactions between center proximity and the mid-level maps (junction and symmetry maps) were very small (see Tables 1, 2, 3), however, the center proximity interactions with low-level and high-level features showed distinct patterns among the models.

The interaction pattern between low-level saliency and center proximity was different in each decomposed deep saliency model. In MSI-Net ($\beta = -0.03$, CI $[-0.04, -0.03]$, $t = -21.6$, $p < .001$), as low-level saliency increased the effect of center proximity decreased (Fig. 3b, bottom-right). In SAM-ResNet ($\beta = 0.04$, CI $[0.03, 0.04]$, $t = 16.72$, $p < .001$), as low-level saliency increased the effect of center proximity increased (Fig. 5b, bottom-right). In DeepGaze II ($\beta = -0.04$, CI $[-0.04, -0.03]$, $t = -30.9$, $p < .001$), a dissociation was observed. That is, as low-level saliency increased, greater center proximity switched from being associated with higher DeepGaze II values to lower DeepGaze II values (Fig. 5b, bottom-right). The interaction pattern between high-level semantic density and center proximity was consistent for MSI-Net and SAM-ResNet

(MSI-Net $\beta = 0.09$, CI $[0.088, 0.095]$, $t = 52.8$, $p < .001$; SAM-ResNet, $\beta = 0.13$, CI $[0.128, 0.137]$, $t = 55.8$, $p < .001$). In both models, as meaning increased center proximity had a greater positive impact on the predicted deep saliency values (Fig. 3b and 5b, top-middle). In comparison, DeepGaze II showed a much smaller interaction between meaning and center proximity ($\beta = 0.027$, CI $[0.025, 0.030]$, $t = 22.2$, $p < .001$). These different interaction patterns with center proximity are likely influenced by both the different model architectures and the different center biases in each deep saliency model.

Discussion

Using deep saliency models to inform cognitive theories of attention requires more than state-of-the-art prediction, it requires an understanding of how that prediction is achieved. Here, we first replicated and extended to active viewing tasks that three of the current best deep saliency models (i.e., MSI-Net, DeepGaze II, and SAM-ResNet) predicted where people looked in real-world scenes. Then, we decomposed the degree to which low-, mid-, and high-level scene information were associated with each deep saliency model. We found that all three models were most strongly associated with features associated with high-level meaning. Low-level image saliency was the second most associated across the three models, while mid-level features were only weakly associated with each deep saliency model.

The strong association between all the deep saliency models we tested and high-level meaning suggests deep saliency models are learning image features that are associated with scene meaning. While MSI-Net, DeepGaze II, and SAM-ResNet each have a unique architecture, training regimen, and loss function, all the models are trained on human scene fixation data. Given previous research indicating scene meaning is one of the strongest predictors of where observers fixate in scenes[45], it follows that deep saliency models would benefit from learning features associated with scene meaning. Therefore, the use of scene fixation data to train deep saliency models may be the common factor that drives each deep saliency model to learn which pre-trained object recognition features are most associated with scene meaning. It is important to note that this does not necessarily mean that deep learning models and human ratings of meaning are equivalent[46]. For example, recent neurocognitive work shows meaning maps produce stronger activation in cortical areas along the ventral visual stream than DeepGaze II[47]. The differences between meaning maps and deep saliency maps are most likely driven by the inherent differences between deep saliency models and human raters. Specifically, deep saliency models have a much simpler neural architecture compared to human raters, and while deep saliency models have a constrained feature set of the visual features stored in VGG-16/VGG-19, human raters likely draw on a much broader set of features including object semantics[30].

The consistent association between deep saliency models and low-level image saliency is also an interesting finding. The deep saliency models each have access to low-level features in the pre-trained VGG-16 and VGG-19 weights of the models. That is, early layers of VGG-16 and VGG-19 both exhibit frequency, orientation, and color selective kernels similar to properties observed in early visual cortex[7,48,49]. Therefore, it is likely that the association we observed between low-level image saliency and each model was driven by the low-level features in the early layers of VGG-16/VGG-19 and the human fixation data during training. Interestingly, while high-level features often override low-level saliency in human observers[25–28,31], it may be that the deep saliency models are learning when low-level features and high-level features are most relevant for predicting where people look in scenes. For example, in all three models we observed a significant interaction between low-level saliency and high-level meaning. And at least in DeepGaze II, the pattern of the interaction seemed consistent with the idea that high-level features can override low-level saliency in scenes. That is, we observed that as a fixated region's meaning value increased, the predicted DeepGaze II values were increasingly unaffected by low-level saliency. Granted we observed the opposite interaction pattern in MSI-Net and SAM-ResNet, so more work will be needed to understand why different deep saliency models show different interaction patterns between low- and high-level scene information. Nonetheless, these results suggest that deep saliency models are learning something about how to balance low- and high-level features, although they seem to be learning different mappings in different deep saliency models.

The mid-level associations with the deep saliency models were relatively weak compared to high-level meaning and low-level saliency. This suggests that local symmetry and junction density, while significant, may only play a supporting role in attentional guidance in scenes. That is, mid-level features help to combine low-level features into higher-level representations[23], but it is these higher-level representations that are used to determine attentional priority. Our mid-level findings using local contour symmetry and junction density complement previous work[9] by examining how these mid-level features are directly associated with fixated deep saliency model values. Finally, it is worth noting that while our current results suggest local symmetry and junction density play marginal roles in MSI-Net, DeepGaze II, and SAM-ResNet, it may simply be that these deep saliency models are using a different kind of mid-level feature representation.

The current work has a number of limitations that would be useful to address in future work to expand our understanding of how deep saliency models predict scene attention. One limitation of the current work is we used active viewing tasks that do not involve a specific target object (i.e., scene memorization and aesthetic judgment). The results will likely be different in a task that involves a search for a specific visual or semantic target (e.g., visual search for a specific object). Another limitation is the current scenes were typical indoor and outdoor scenes, without semantically inconsistent objects. So it will be important in

future work to examine whether similar patterns of association hold for scenes that contain object-scene inconsistency[50–53]. Finally, we only looked at two possible mid-level features, so it would be useful in future work to test other candidate mid-level features. For example, one could investigate the intermediate layers of VGG-16/VGG-19, or other proposed mid-level feature representations such as texforms[23]. Fortunately, the general approach introduced here can easily be extended to examine other candidate mid-level feature representations.

While deep learning models provide state-of-the-art scene fixation prediction, insights they might provide for cognitive theories of attention have been limited. In order for deep saliency models to inform cognitive theories of gaze guidance in scenes, we must find ways to understand the feature mapping these models are learning from the human data. Here, we have shown one approach to decomposing the performance of deep saliency models by using maps that reflect a wide range of processing levels ranging from pre-semantic, low-level image saliency to high-level meaning. We found that all three deep saliency models were most strongly associated with high-level meaning, followed by low-level image saliency, and mid-level contour symmetry and junctions. This work supports the importance of meaning in attentional control in real-world scenes, and also provides a general approach for quantifying the association and interaction patterns between overt attention, any feature map, and any deep saliency model.

Methods

Participants

University of California, Davis undergraduate students with normal or corrected-to-normal vision participated in the eye tracking (N=114) and meaning rating (N=408) studies in exchange for course credit. All participants were naive concerning the purposes of the experiment and provided verbal or written informed consent as approved by the University of California, Davis Institutional Review Board.

Stimuli

Participants in the eye tracking study viewed 100 real-world scene images. The 100 scenes were chosen to represent 100 unique scene categories (e.g., kitchen, park), where half of the images were indoor scenes and half were outdoor. Each participant in the meaning rating study viewed and rated 300 isolated, random small regions taken from the set of 100 scenes.

Apparatus

Eye movements were recorded using an EyeLink 1000+ tower-mount eye tracker (spatial resolution 0.01°) sampling at 1000 Hz[54]. Participants sat 85 cm away from a 21" monitor and viewed scenes that subtended approximately 27° x 20° of visual angle. Head movements were minimized using a chin and forehead rest. Although viewing was binocular, eye movements were recorded from the right eye. The display presentation was controlled with SR Research Experiment Builder software[55].

Eye tracking calibration and data quality

A 13-point calibration procedure was performed at the start of each session to map eye position to screen coordinates. Successful calibration required an average error of less than 0.49° and a maximum error of less than 0.99°. Fixations and saccades were segmented with EyeLink's standard algorithm using velocity and acceleration thresholds ($30/s$ and $9500°/s^2$). A drift correction was performed before each trial and recalibrations were performed as needed. The recorded data were examined for data artifacts from excessive blinking or calibration loss based on mean percent signal across trials[56]. Fourteen subjects with less than 75% signal were removed, leaving 100 subjects that were tracked well (signal mean=92.1%, SD=5.31%).

Eye Tracking Tasks and Procedure

Each participant (N=100) viewed 100 scenes for 12 seconds each while we recorded their eye movements. Each trial began with fixation on a cross at the center of the display for 300 ms. For half the scenes, participants were instructed to memorize each scene in preparation for a later memory test. For the other half of the scenes, participants were instructed to indicate how much they liked each scene on a 1-3 scale using a keyboard press following the 12 second scene presentation. The scene set and presentation order of the two tasks were counterbalanced across subjects. This procedure produced a large eye movement dataset that contained 334,725 fixations, with an average of 3347 fixations per subject.

Deep Saliency Models

We compared 3 of the best performing deep saliency models on the MIT saliency benchmark[1]: the multi-scale information network (MSI-Net)[36], DeepGaze II[37], and the saliency attentive model (SAM-ResNet)[38]. Each deep saliency model takes an image as input and produces a predicted saliency map as output. All of the deep saliency models are trained on human data in the form of fixation and/or mouse-contingent density maps that reflect where humans focus their attention in scenes. The model weights are fixed following training, and then the models are evaluated on new scenes and fixation data. MSI-Net, DeepGaze II,

and SAM-ResNet each have distinct network architectures, training regimens, center bias priors, and loss functions which are worth considering.

MSI-Net

MSI-Net consists of three main components, a feature network, a spatial pooling network, and a readout network[36]. MSI-Net uses the pre-trained weights from the VGG-16 network[7] without the feature downsampling in the last two max-pooling layers[36]. The VGG-16 network is a deep convolutional network with 16 layers and was trained on both the *ImageNet* object classification[57] and the *Places2* scene classification datasets[58]. The activations from the VGG-16 network then feed into a spatial pooling module called the *Atrous Spatial Pyramid Pooling* (ASPP) module[59]. The ASPP module of MSI-Net has several convolutional layers which combine feature information at multiple spatial scales including a global scale to capture global scene context which has been shown to be helpful in predicting where people look in scenes[60]. Finally, the readout network contains 6 layers that include convolutional and upsampling layers and a blur. The ASPP and readout network were trained on the SALICON dataset[8]. MSI-Net prediction is optimized using the Kullback–Leibler divergence which measures the distance between the target and model estimated distributions. The MSI-Net predicted saliency maps reflect the predicted probability distribution of fixations for each scene image (Fig. 1b).

DeepGaze II

DeepGaze II also consists of three main components, a feature network, a readout network, and an explicit (i.e., non-learned) center bias[37]. The feature network consists of the pre-trained weights from the VGG-19 network[7] without the fully connected layers. The VGG-19 network is a deep convolutional network with 19 layers that is trained on more than a million images to recognize 1000 different object categories from the *ImageNet* database[57]. In DeepGaze II, the VGG-19 feature network is fixed and the readout network is the only portion of the model that is trained to perform saliency prediction. The readout network consists of 4 layers that are trained on the SALICON[8] and MIT1003[61] datasets to predict human saliency data using the pre-trained VGG-19 features[8]. The DeepGaze II model maximizes log-likelihood and expresses saliency as probability density with blur. Finally, DeepGaze II applies a center bias to capture the tendency for observers to look more centrally in scenes[39,40]. The DeepGaze II maps reflect the predicted probability distribution of fixations for each scene image (Fig. 1c).

SAM-ResNet

SAM-ResNet is composed of a dilated feature network, an attentive convolutional network, and a learned set of Gaussian priors for center bias[38]. SAM-ResNet modifies the ResNet-50 network[62] using a dilation technique to reduce the amount of input image rescaling that is detrimental to saliency prediction[63]. The ResNet-50 network is a deep convolutional network with 50 layers that is trained on the *ImageNet* object classification dataset[57]. The dilated version of the ResNet-50 feature network provides the features that then feed into the attentive convolutional network. The attentive convolutional network is a recurrent long short-term memory network (LSTM,[64]) that is used to refine the most salient regions of the input regions over multiple sequential iterations. It is worth noting that this recurrent model module is fundamentally different compared to the pure feedforward MSI-Net and DeepGaze II model architectures[38]. Finally, SAM-ResNet learns a set of Guassian priors to account for observer center bias[39,40]. SAM-ResNet is trained on the SALICON dataset[8] and uses a linear combination of multiple saliency benchmark metrics (i.e., normalized scanpath saliency, linear correlation, and Kullback–Leibler divergence,[1]) as its loss function during training. The SAM-ResNet predicted saliency maps reflect the predicted probability distribution of fixations for each scene image (Fig. 1d).

Feature Maps

Low-level features: Image saliency map

Low-level scene features were represented using the Itti and Koch model with blur with default settings[3,42,65]. Similar to other image-based saliency models, the Itti and Koch model is derived from contrasts in low-level image features including color, intensity, and edge orientation at multiple spatial scales. An image saliency map was generated for each scene stimulus and reflects the predicted fixation density for each scene based on low-level, pre-semantic image features.

Mid-level features: Symmetry and junction maps

Mid-level scene features were represented by two different types of maps: symmetry maps and junction maps. The symmetry and junction maps were both computed from a line drawing of each scene. The line drawings (Fig. 6b) were extracted using an automated line drawing extraction algorithm (logical/linear operators,[43,66]). Then, using the contours from each scene line drawing, the symmetry (Fig. 6c) and junction (Fig. 6d) maps were computed. Each scene symmetry map reflects the degree of local ribbon symmetry of contours in the scene line drawing[43,44]. Ribbon symmetry measures the degree to which pairs of scene contours exhibit constant separation (i.e., local parallelism) along their medial axis[43,44]. Each scene junction map shows the density of points where at least two separate scene contours intersect each other[67].

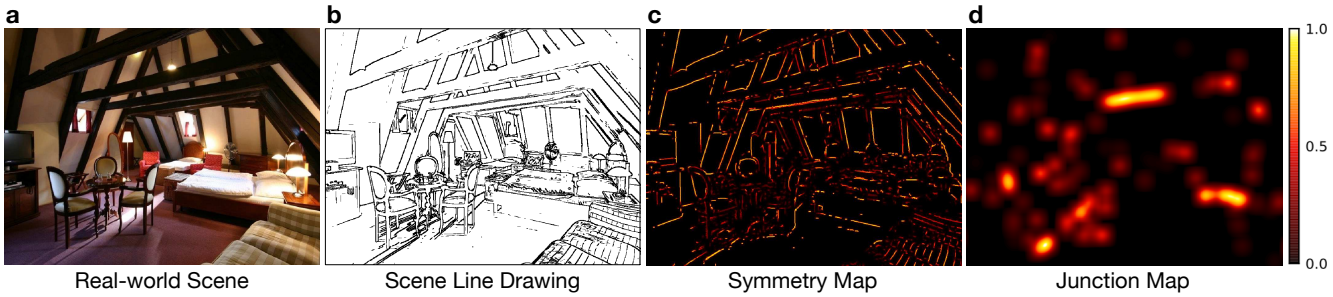

Figure 6. *Scene, line drawing, and its corresponding symmetry and junction maps.* Each scene (**a**) was first converted to a line drawing (**b**). Then, from the line drawing, local symmetry (**c**) and junction density were computed (**d**). The symmetry and junction maps served as mid-level feature maps in our analyses.

High-level features: Meaning Map

Meaning maps were generated as a representation of the spatial distribution of high-level, semantic density ([26,27]; see https://osf.io/654uh/ for code and task instructions). Meaning maps were created for each scene by cutting the scene into a dense array of overlapping circular patches at a fine spatial scale (300 patches with a diameter of 87 pixels) and coarse spatial scale (108 patches with a diameter of 207 pixels). Each rater (N=408) then provided ratings of 300 random scene patches based on how informative or recognizable they thought they were on a 6-point Likert scale[24,26]. Patches were presented in random order and without scene context, so ratings were based on context-independent judgments. Each patch was rated by 3 unique raters. A meaning map (Fig. 1f) was generated for each scene by averaging the rating data at each spatial scale separately, then averaging the spatial scale maps together, and then smoothing the grand average rating map with a Gaussian filter (i.e., Matlab 'imgaussfilt' with $\sigma = 10$, $FWHM$=23 px).

Center Proximity Map

In addition to the low-, mid-, and high-level feature maps, we also generated a center proximity map that served as a global representation of how far each location in the scene was from the scene center. Specifically, the center proximity map measured the inverted Euclidean distance from the center pixel of the scene to all other pixels in the scene image (Fig. 1e). The center proximity map[30] was used to explicitly control for the general bias for observers to look more centrally than peripherally in scenes, independent of the underlying scene content[39,40].

Statistical Models

Fixated and non-fixated scene locations

We modeled the association between the eye movement data and each deep saliency model by comparing where each subject looked in each scene to where they did not look[30,41]. Specifically, for each region a subject fixated, we computed the mean value for each deep saliency model (Fig. 1b, 1c, 1d) and the center proximity map (Fig. 1e) by taking the average over a 3° window around each fixation (Fig. 1a, neon green locations). To represent the model and center proximity values that were not associated with overt attention, for each individual subject, we randomly sampled an equal number of scene locations where each subject did not look in each scene they viewed (Fig. 1a, cyan locations). The only constraint for the random sampling of the non-fixated scene regions was that the non-fixated 3° windows could not overlap with any of the fixated 3° windows. This procedure was performed separately for each individual scene viewed by each individual subject.

Fixation location: General linear mixed effects models

We applied a general linear mixed effects (GLME) logit model to examine how well each deep saliency model accounted for the eye movement data using the *lme4* package[68] in R[69]. We used a mixed effects modeling approach because it does not require aggregating the eye movement data at the subject or scene-level like ANOVA or map-level correlations. Instead, both subject and scene could be explicitly modeled as random effects. The GLME approach allowed us to control for center bias by including the center proximity (Fig. 1e) of each fixated and non-fixated region as both a fixed effect and as an interaction term with the deep saliency model values. Specifically, whether a region was fixated (1) or not fixated (0) was modeled as a function of the fixed effects of each deep saliency map value (i.e., MSI-Net, DeepGaze II, or SAM-ResNet), center proximity value, and the deep saliency model by center proximity interaction. Subject and scene were treated as random intercepts. Since we are interested in how well each deep saliency model performs generally, regardless of task, the memorization and aesthetic

judgment data were combined in all models. To compare the performance of the 3 different deep saliency models, a GLME model was fit separately for each deep saliency model (see Fig. 2).

Quantifying the roles of low-, mid-, and high-level features: Linear mixed effects models

In addition to how well the deep saliency models acccounted for overt attention, we also quantified the associations between low-, mid-, and high-level features and each deep saliency mode by fitting a linear mixed effects (LME) model to each deep saliency model. In each LME model, the fixated deep saliency model values (i.e., MSI-Net, DeepGaze II, or SAM-ResNet) were modeled as a function of the fixed effects of center proximity (bias), Itti and Koch image saliency (low-level), symmetry and junction (mid-level), and meaning (high-level). Given the known strong effect of center bias[39,40] we included center proximity as an interaction term with all other feature maps. Finally, since high-level and low-level features are known to be associated with each other[26,32], we included a low-level by high-level feature interaction term (i.e., Itti & Koch X Meaning) in each deep saliency LME model. Conceptually, these LME models for each deep saliency model (MSI-Net, Fig. 3; DeepGaze II, Fig. 4; SAM-ResNet, Fig. 5) estimate the degree to which the various feature maps (i.e., low-, mid-, and high-level) are related to the respective deep saliency model output.

References

1. Bylinskii, Z. *et al.* MIT Saliency Benchmark. http://saliency.mit.edu/ (2012).
2. Borji, A., Sihite, D. N. & Itti, L. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Process.* **22**, 55–69 (2013).
3. Harel, J., Koch, C. & Perona, P. Graph-based visual saliency. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, 545–552 (MIT Press, Cambridge, MA, USA, 2006).
4. Itti, L. & Koch, C. Computational modeling of visual attention. *Nat. Rev. Neurosci.* **2**, 194–203 (2001).
5. Koch, C. & Ullman, U. Shifts in selective visual attention: Towards a underlying neural circuitry. *Hum. Neurobiol.* **4**, 219–227 (1985).
6. Parkhurst, D., Law, K. & Niebur, E. Modeling the role of salience in the allocation of overt visual attention. *Vis. Res.* **42**, 102–123 (2002).
7. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *CoRR* **abs/1409.1556** (2015).
8. Jiang, M., Huang, S., Duan, J. & Zhao, Q. Salicon: Saliency in context. *2015 IEEE Conf. on Comput. Vis. Pattern Recognit. (CVPR)* 1072–1080 (2015).
9. Damiano, C., Wilder, J. D. & Walther, D. B. Mid-level feature contributions to category-specific gaze guidance. *Attention, Perception, & Psychophys.* **81**, 35–46 (2019).
10. Treisman, A. & Gelade, G. A feature integration theory of attention. *Cogn. Psychol.* **12**, 97–136 (1980).
11. Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. *Annu. review neuroscience* **18**, 193–222 (1995).
12. Wolfe, J. M. & Horowitz, T. S. Five factors that guide attention in visual search. *Nat. Hum. Behav.* **1**, 1–8 (2017).
13. Allman, J., Miezin, F. M. & McGuinness, E. Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annu. review neuroscience* **8**, 407–30 (1985).
14. Desimone, R., Schein, S. J., Moran, J. P. & Ungerleider, L. G. Contour, color and shape analysis beyond the striate cortex. *Vis. Res.* **25**, 441–452 (1985).
15. Knierim, J. J. & Essen, D. C. V. Neuronal responses to static texture patterns in area v1 of the alert macaque monkey. *J. neurophysiology* **67 4**, 961–80 (1992).
16. Garcia-Diaz, A., Leborán, V., Fdez-Vidal, X. R. & Pardo, X. On the relationship between optical variability, visual saliency, and eye fixations: a computational approach. *J. vision* **12 6**, 17 (2012).
17. Riche, N. *et al.* Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Process. Image Commun.* **28**, 642–658, DOI: https://doi.org/10.1016/j.image.2013.03.009 (2013).
18. Bruce, N. D. & Tsotsos, J. K. Saliency, attention, and visual search: An information theoretic approach. *J. Vis.* **9**, 1–24 (2009).

19. Koffka, K. *Principles of Gestalt Psychology* (Harcourt, Brace, New York, 1935).

20. Wertheimer, M. Laws of organization in perceptual forms. In Ellis, W. B. (ed.) *A Sourcebook of Gestalt Psychology*, 71–88 (Harcourt: Brace and Company, 1938).

21. Biederman, I. Recognition-by-components: a theory of human image understanding. *Psychol. review* **94 2**, 115–147 (1987).

22. Wagemans, J. *et al.* A century of gestalt psychology in visual perception: I. perceptual grouping and figure-ground organization. *Psychol. bulletin* **138 6**, 1172–217 (2012).

23. Long, B., Yu, C. & Konkle, T. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proc. Natl. Acad. Sci.* **115**, 9015–9024 (2018).

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

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

26. Henderson, J. M. & Hayes, T. R. Meaning-based guidance of attention in scenes rereveal by meaning maps. *Nat. Hum. Behav.* **1**, 743–747 (2017).

27. Henderson, J. M. & Hayes, T. R. Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *J. Vis.* **18**, 1–18 (2018).

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

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

30. Hayes, T. R. & Henderson, J. M. Looking for semantic similarity: What a vector space model of semantics can tell us about attention in real-world scenes. *Psychol. Sci.* 1–7 (In Press).

31. Hart, B. M., Schmidt, H., Roth, C. & Einhäuser, W. Fixations on objects in natural scenes: dissociating importance from salience. *Front. Psychol.* **4**, 1–9 (2013).

32. Hayes, T. R. & Henderson, J. M. Scene semantics involuntarily guide attention during visual search. *Psychon. Bull. Rev.* 1–7, DOI: 10.3758/s13423-019-01642-5 (2019).

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

34. Henderson, J. M., Hayes, T. R., Rehrig, G. & Ferreira, F. Meaning guides attention during real-world scene description. *Sci. Reports* **8**, 1–9 (2018).

35. Peacock, C. E., Hayes, T. R. & Henderson, J. M. Meaning guides attention during scene viewing even when it is irrelevant. *Attention, Perception, & Psychophys.* **81**, 20–34 (2019).

36. Kroner, A., Senden, M., Driessens, K. & Goebel, R. Contextual encoder-decoder network for visual saliency prediction. *Neural networks : official journal Int. Neural Netw. Soc.* **129**, 261–270 (2020).

37. Kümmerer, M., Wallis, T. S. A. & Bethge, M. Deepgaze II: reading fixations from deep features trained on object recognition. *CoRR* **abs/1610.01563** (2016). 1610.01563.

38. Cornia, M., Baraldi, L., Serra, G. & Cucchiara, R. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Process.* **27**, 5142–5154 (2018).

39. Tatler, B. W. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *J. Vis.* **7**, 1–17 (2007).

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

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

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

43. Rezanejad, M. *et al.* Scene categorization from contours: Medial axis based salience measures. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

44. Wilder, J. *et al.* Local contour symmetry facilitates scene categorization. *Cognition* **182**, 307–317, DOI: 10.1016/j.cognition.2018.09.014 (2019).

45. Henderson, J. M., Hayes, T. R., Peacock, C. E. & Rehrig, G. Meaning and attentional guidance in scenes: A review of the meaning map approach. *Vision* **2**, 1–10 (2019).

46. Henderson, J. M., Hayes, T. R., Peacock, C. E. & Rehrig, G. Meaning maps capture the density of local semantic features in scenes: A reply to Pedziwiatr, Kummerer, Wallis, Bethge & Teufel (2021). *Cognition* 1–16 (Under Review).

47. Henderson, J. M., Goold, J. E., Hayes, T. R. & Choi, W. Neural Correlates of Image Salience and Semantic Content during Active Scene Viewing: Evidence from Fixation-Related fMRI. 1–37 (Under Review).

48. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *NIPS* 1097–1105 (2012).

49. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. In Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T. (eds.) *Computer Vision – ECCV 2014*, 818–833 (Springer International Publishing, Cham, 2014).

50. Loftus, G. R. & Mackworth, N. H. Cognitive determinants of fixation location during picture viewing. *J. Exp. Psychol.* **4**, 565–572 (1978).

51. Henderson, J. M., Weeks, P. A. & Hollingworth, A. The effects of semantic consistency on eye movements during complex scene viewing. *J. Exp. Psychol. Hum. Percept. Perform.* **25**, 210–228 (1999).

52. Brockmole, J. R. & Henderson, J. M. Prioritizing new objects for eye fixation in real-world scenes: Effects of object-scene consistency. *Vis. Cogn.* **16**, 375–390 (2008).

53. Võ, M. L. H. & Henderson, J. M. Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *J. Vis.* **9**, 1–15 (2009).

54. SR Research. *EyeLink 1000 User's Manual, Version 1.5.2* (SR Research Ltd., Mississauga, ON, 2010).

55. SR Research. *Experiment Builder User's Manual* (SR Research Ltd., Mississauga, ON, 2010).

56. Holmqvist, K., Nyström, R., M.and Andersson, Dewhurst, R., Jorodzka, H. & van de Weijer, J. *Eye Tracking: A comprehensive guide to methods and measures* (Oxford University Press, 2015).

57. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. *2009 IEEE Conf. on Comput. Vis. Pattern Recognit.* 248–255 (2009).

58. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A. & Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis Mach. Intell.* **40**, 1452–1464 (2018).

59. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis Mach. Intell.* **40**, 834–848 (2018).

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

61. Judd, T., Ehinger, K. A., Durand, F. & A., T. Learning to predict where humans look. *2009 IEEE 12th Int. Conf. on Comput. Vis.* 2106–2113 (2009).

62. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *2016 IEEE Conf. on Comput. Vis. Pattern Recognit. (CVPR)* 770–778 (2016).

63. Yu, F. & Koltun, V. Multi-scale context aggregation by dilated convolutions. *CoRR* **abs/1511.07122** (2016).

64. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).

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

66. Iverson, L. A. & Zucker, S. W. Logical/linear operators for image curves. *IEEE Transactions on Pattern Analysis Mach. Intell.* **17**, 982–996, DOI: 10.1109/34.464562 (1995).

67. Walther, D. B. & Shen, D. Nonaccidental properties underlie human categorization of complex natural scenes. *Psychol. Sci.* **25**, 851–860 (2014).

68. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48, DOI: 10.18637/jss.v067.i01 (2015).

69. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2017).

Acknowledgements

This research was supported by the National Eye Institute of the National Institutes of Health, under award number R01EY027792. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors declare no competing financial interests.

Author contributions statement

T.R.H. and J.M.H. conceived the study, T.R.H conducted the experiments and analyzed the data. T.R.H. drafted the manuscript and J.M.H. revised the manuscript.

Figures

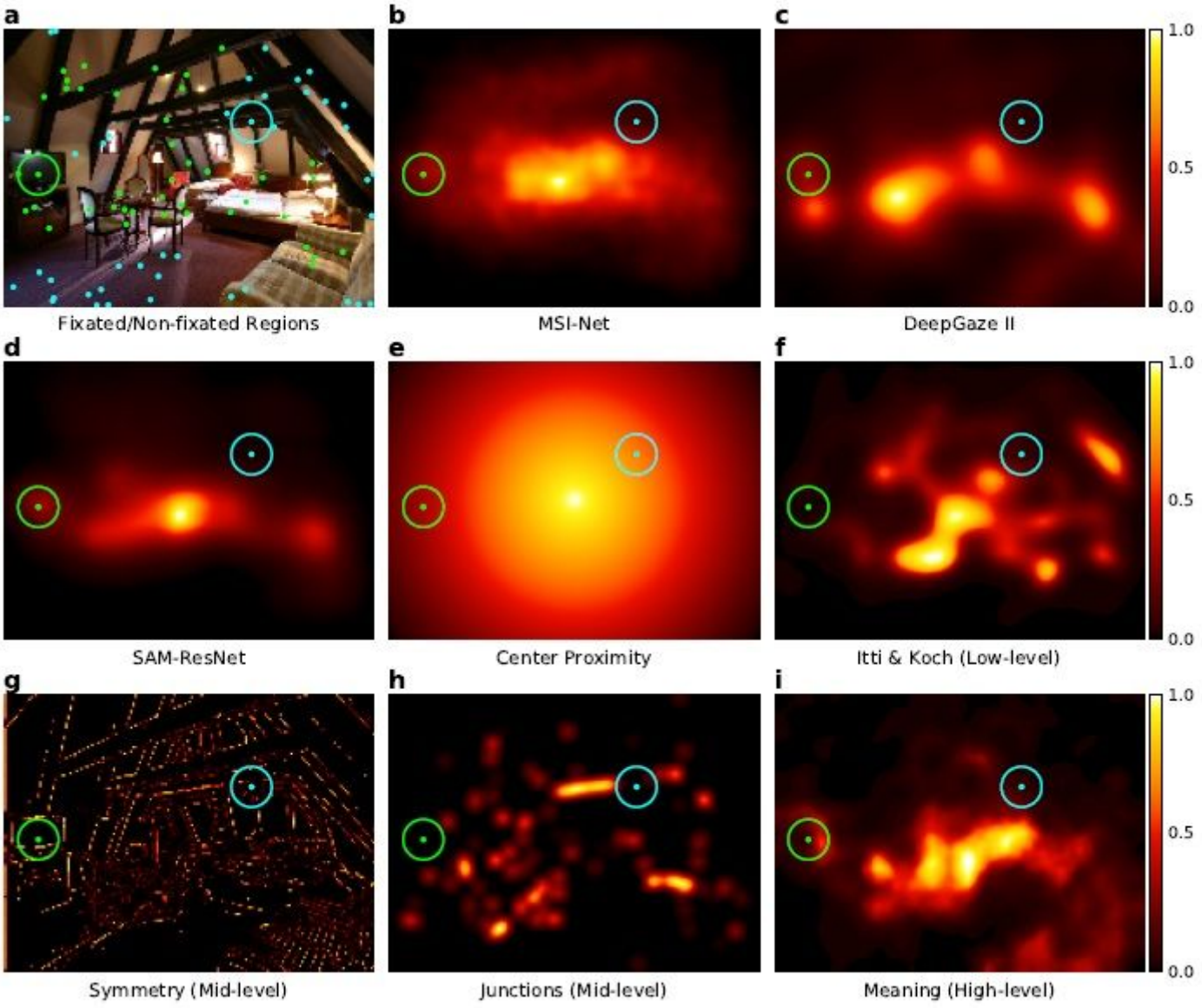

Figure 1

Scene with the fixated and non-fixated regions for a single subject, and the corresponding deep saliency and feature maps. a The green dots show the fixation locations for a single viewer and the cyan dots indicate randomly sampled non-fixated regions that represent where this subject did not look in this scene. Together these locations provide an account of the regions in this scene that did and did not capture this subject's attention. Each fixated and non-fixated location was then used to compute a mean model value for each deep saliency model map (b, c, d) and feature map (e, f, g, h, i) across a 3o window (shown as circles around one example fixated and non-fixated location).

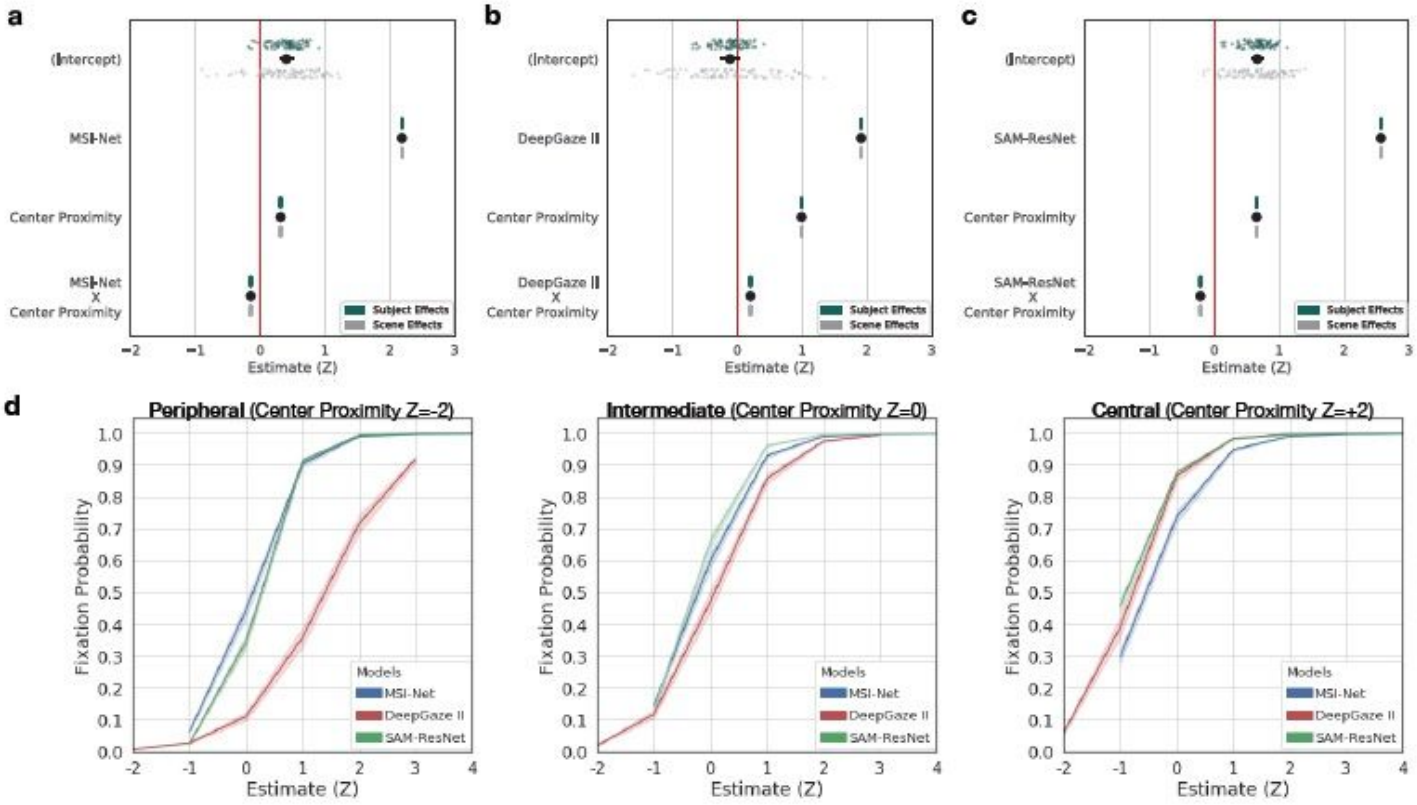

Figure 2

Deep saliency model general linear mixed effects model results. Whether a scene region was fixated or not was modeled as a function of the deep saliency model value, center proximity value, and their interaction as fixed effects (a MSI-Net; b DeepGaze II; c SAM-ResNet). The black dots with lines show the fixed effect estimates and their 95% confidence intervals. Subject (green dots) and scene (grey dots) were both accounted for in the model as random intercepts. d A line plot of the interaction between center proximity (panels) and each deep saliency model (colored lines) as a function of fixation probability. All error bands reflect 95% confidence intervals.

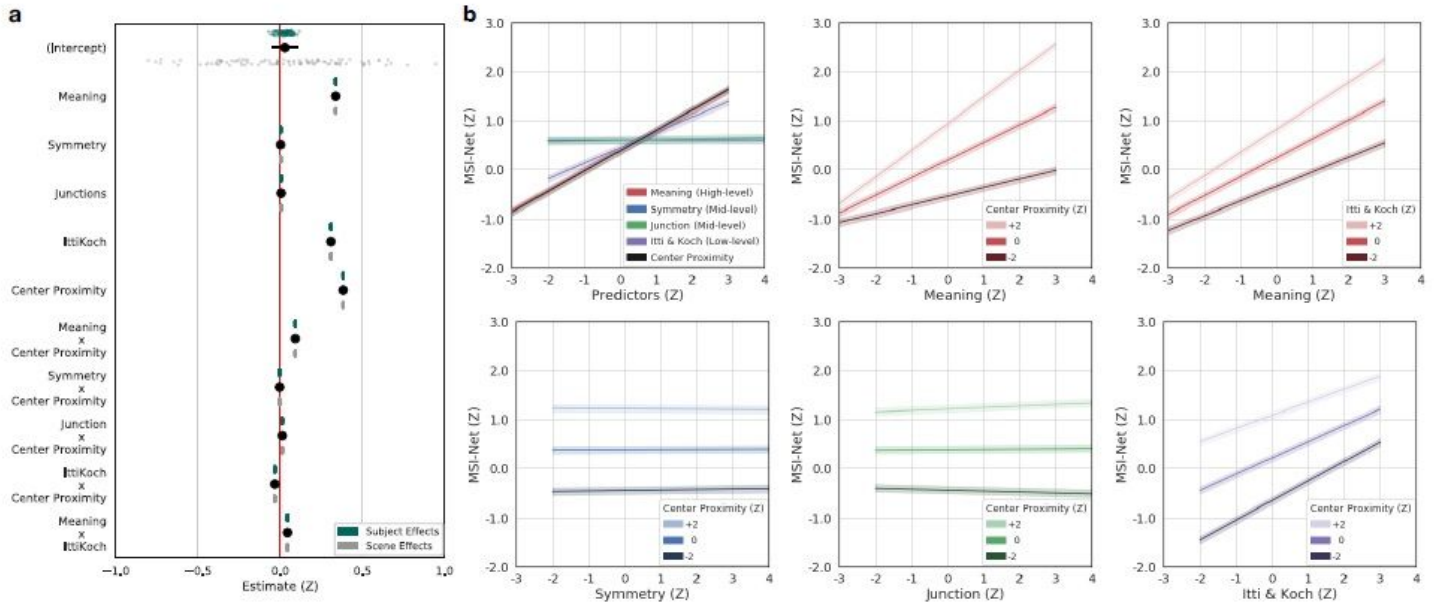

Figure 3

MSI-Net linear mixed effects model, marginal effects, and interactions. a Fixated MSI-Net values as a function of low-, mid-, and high-level features and interactions. The blackdots with lines show the fixed effect estimates and their 95% confidence intervals. Subject (greendots) and scene (grey dots) were both accounted for in the model as random intercepts. b Line plots of all model marginal effects and all model interactions. All error bands reflect 95% confidence intervals.

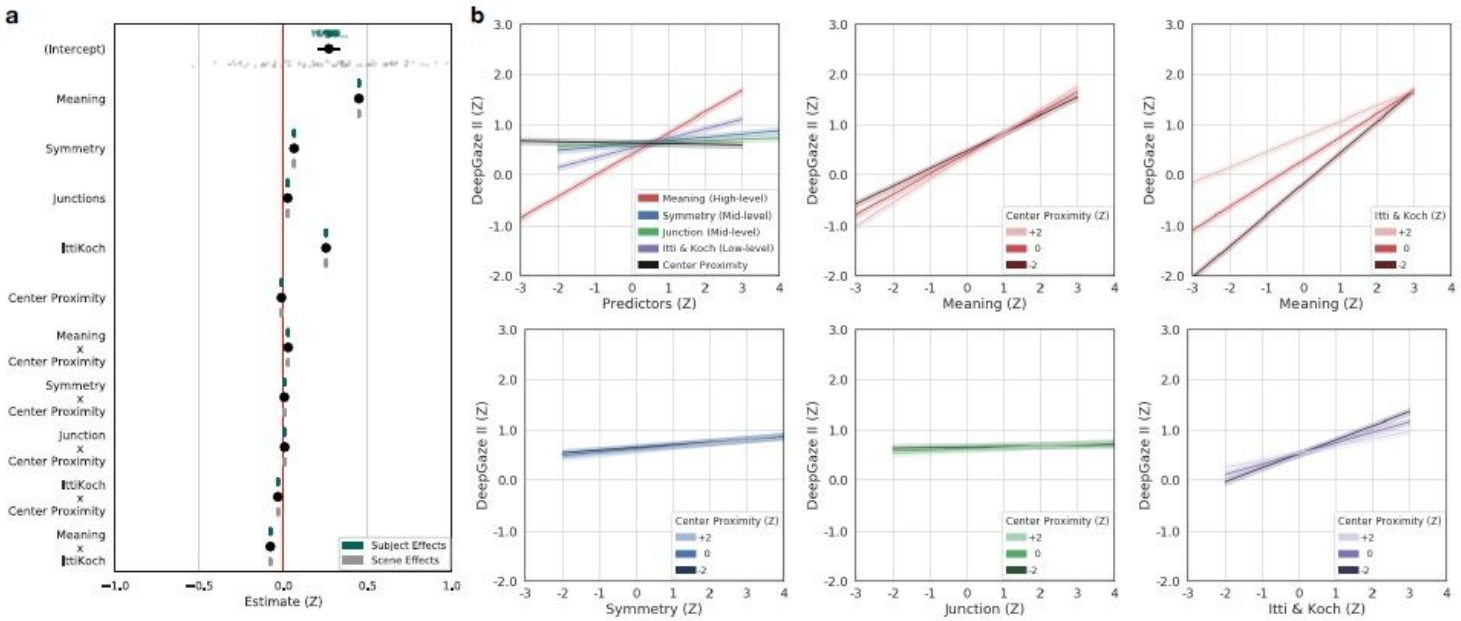

Figure 4

DeepGaze II LME model, marginal effects, and interactions. a Fixated DeepGaze II values as a function of low-, mid-, and high-level features and interactions. The blackdots with lines show the fixed effect

estimates and their 95% confidence intervals. Subject (greendots) and scene (grey dots) were both accounted for in the model as random intercepts. b Line plots of all model marginal effects and all model interactions. All error bands reflect 95% confidence intervals.

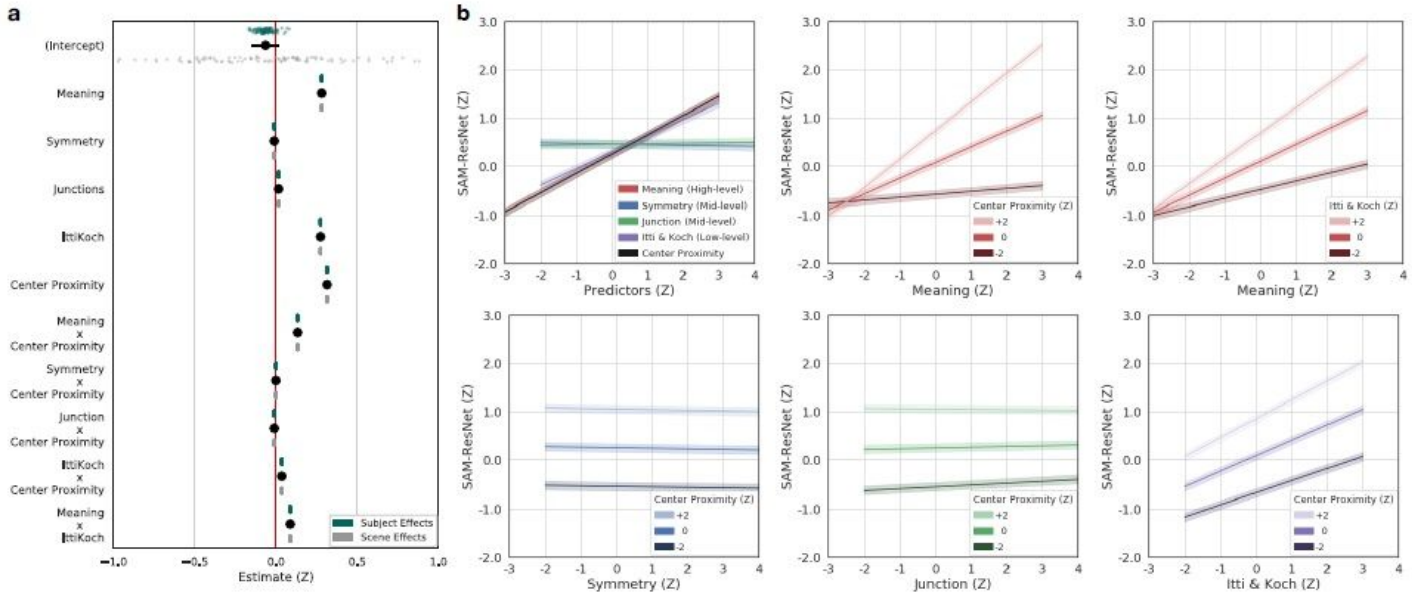

Figure 5

SAM-ResNet LME model, marginal effects, and interactions. a Fixated SAM-ResNet values as a function of low-, mid-, and high-level features and interactions. The blackdots with lines show the fixed effect estimates and their 95% confidence intervals. Subject (greendots) and scene (grey dots) were both accounted for in the model as random intercepts. b Line plots of all model marginal effects and all model interactions. All error bands reflect 95% confidence intervals.

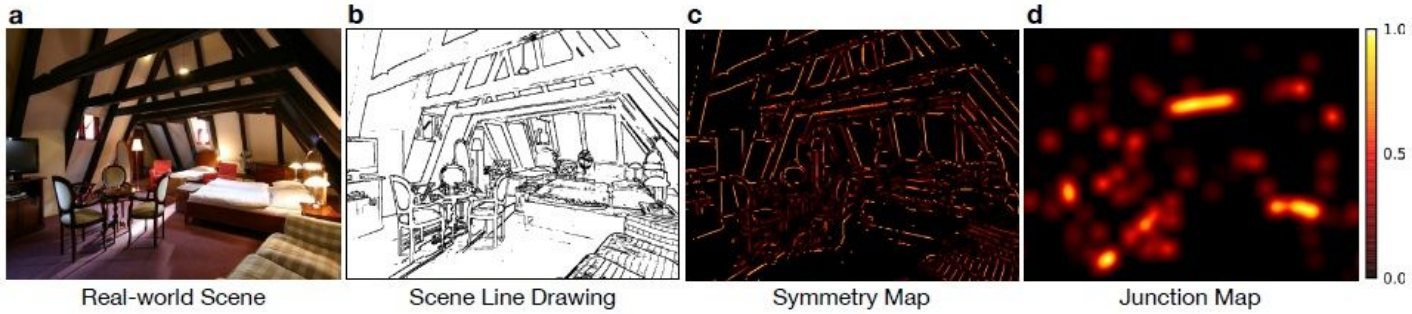

Figure 6

Scene, line drawing, and its corresponding symmetry and junction maps. Each scene (a) was first converted to a line drawing (b). Then, from the line drawing, local symmetry (c) and junction density were computed (d). The symmetry and junction maps served as mid-level feature maps in our analyses.