DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling Supplementary Material

As mentioned in the main text, our initial experiments involved layer selection experiments (10 experiments with 10 different configurations) followed by 20 repeated experiments for initialization selection (figure 1). We also conducted a detailed evaluation on MIT1003 on all backbones and saliency metrics (table 1). Our combinatoric experiments were firstly inspired from highlighting how different backbones perform significantly different on a per-sample basis (figure 2). Towards the making of our ensembles, we first tried different weights and found that performance consistently peaks when different backbones get an equal say on the prediction (figure 3). Finally, we conducted a principled qualitative analysis using samples whose predictions are maximally different across the backbones of our ensemble (figure 4).

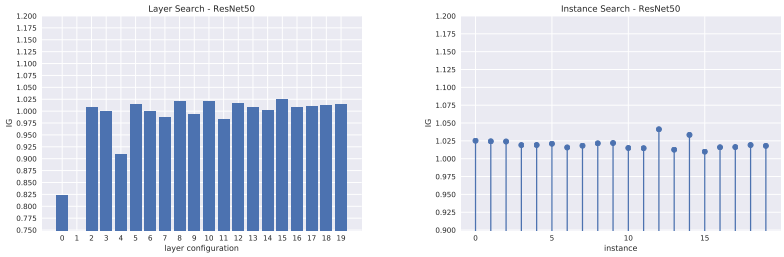

Figure 1: ResNet50 layer search (top image) reflects our experiments that involved trying out layers from ResNet50's final convolutional blocks as features. In the case of instance search (bottom image) we simply pick the top performing layer configuration and repeat the same experiment with different seeds (hence different initialization). We can see that the fluctuations between different instances are just as high as the ones we see among the top 5 layer configurations.

Table 1: Evaluation on all models on MIT1003. To assure the robustness of our metrics, we calculate performance in each metric for 20 instances per model, then take the average

Backbone	IG ↑	AUC ↑	sAUC ↑	NSS ↑	CC ↑	KLDiv ↓
densenet201	**1.0377**	**0.8892**	**0.7876**	2.5994	**0.7736**	**0.5156**
resnext50	1.0368	0.8886	0.7854	2.6354	0.7731	0.5214
efficientnet	1.0326	0.8890	0.7870	2.6213	0.7704	0.5237
shapenetC	1.0278	0.8878	0.7848	**2.6380**	0.7716	0.5263
resnet50	1.0201	0.8874	0.7834	2.6141	0.7657	0.5318
resnet101	1.0045	0.8866	0.7816	2.5909	0.7631	0.5389
vgg19	0.9483	0.8838	0.7747	2.5457	0.7486	0.5653
vgg11	0.9035	0.8803	0.7681	2.4905	0.7346	0.5896
alexnet	0.8046	0.8736	0.7554	2.3073	0.6983	0.6482

Table 2: Performance of DeepGaze IIE on the SALICON test set. For this version of DeepGaze IIE, we average the individual models after pretraining on the SALICON training dataset, i.e. without finetuning on MIT1003. The SALICON competition does not support proper evaluation of probabilistic models, but only of classic saliency maps. Therefore all reported scores are for saliency maps optimal for NSS (i.e. predicted fixation densities), except for sAUC, for which we used the correct saliency maps for sAUC (i.e., predicted fixation density divided by the average of the predicted fixation densities for all other images).

Model	sAUC	IG	NSS	CC	AUC	SIM	KL
DeepGaze IIE	0.767	0.766	1.996	0.872	0.869	0.733	0.285

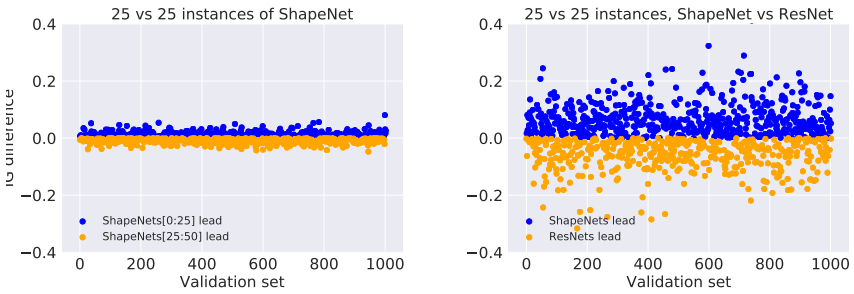

Figure 2: Per-image performance variance in between different models. Each point on the X axis corresponds to an image from MIT1003 while the Y axis is the information gain *difference* between the two groups of models, meaning information gain was calculated and averaged across one group of models then subtracted between the two. Thus, the different colors signify which of the two groups is leading in the corresponding sample. On the left plot, we compare 50 instances of ShapeNetC in groups of 25 and find that even with the exact same architecture, the standard deviation is a non-marginal value of 0.015. However, when we compared groups of ShapeNetC to ResNet50 (right plot) we found a significant standard deviation of 0.086 in their per-image information gain difference.

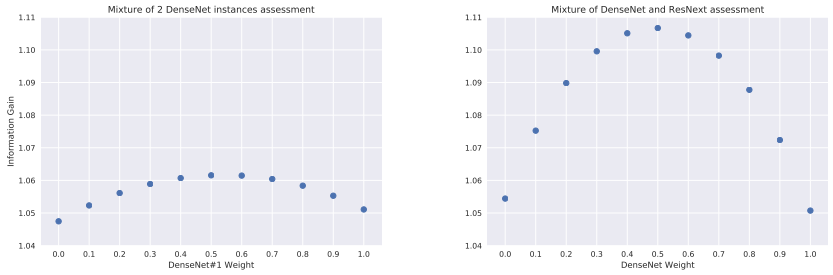

Figure 3: Mixtures of models with varying weights. We show the performance when using a mixture of two models with varying mixing coefficients, so that at 0 we see the individual performance of one instance, at 1 that of the other instance while at 0.5 both have equal say at the final density. The left figure shows performances of average densities from instances that use DenseNet-201 as a backbone and is indicative of intra-model complementarity while in the right figure it's instances from two distinct distinct backbones (ResNext50 and DenseNet201) and indicative of inter-model complementarity. Even when mixing instances of the exact same model there is a boost in performance that peaks at the point where each model has equal weight; however, we reach a much higher performance when mixing instances of different models. We empirically found this to be true when combining other models presented in this paper as well.

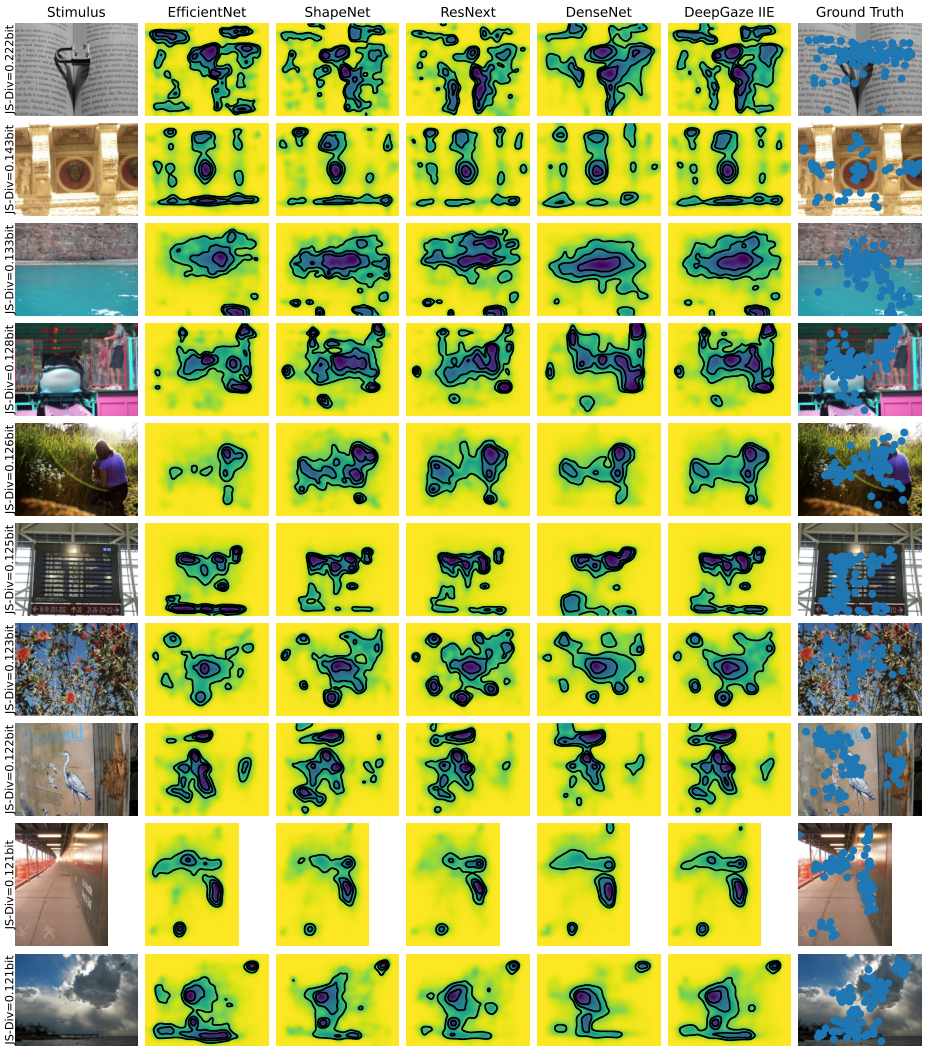

Figure 4: We predict the fixation densities from different models using samples of the MIT1003 dataset. To select samples where our models are qualitatively different, we compute the Jensen-Shannon divergence (JS-Div) per image amongst our top four models (not including the mixture DSRE), using mixture of 20 instances per model, thus removing any noise caused by intra-model variability. These are the top-10 images in terms of maximal disagreement and are displayed top to bottom with respect to their JS-Div, the maximum being 0.222 bits corresponding to the top row image.