SUPPLEMENTAL MATERIAL: Where should saliency models look next?

Zoya Bylinskii[1], Adrià Recasens[1], Ali Borji[2],
Aude Oliva[1], Antonio Torralba[1], and Frédo Durand[1]

[1] Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
{zoya,recasens,oliva,torralba,fredo}@mit.edu
[2] Center for Research in Computer Vision
University of Central Florida
aborji@crcv.ucf.edu

Evaluating progress

Top performers on the MIT Benchmark

In Fig. 1 we include the performances of the top four neural network models evaluated on the MIT300 dataset (as of March 2016), along with the top three non neural network models, and three traditional bottom-up approaches that are commonly used for saliency comparisons. The metrics reported are ones evaluated on the MIT Saliency Benchmark [1][3], supplemented with information gain (as recommended in [2]), and discussed in detail in Bylinskii et al. [3]. From this plot we can see that the top neural network model scores are significantly better than prior model scores, and these are the models we use for our main analyses to quantify the remaining gap to human performance. At the same time, we see that the AUC metrics have begun to saturate and are not as informative at continuing to track model performances.

In Table 1 the top neural network model (DeepFix) is compared to the top non neural network model (BMS) on the CAT2000 dataset, the second benchmark dataset on the MIT Saliency Benchmark. While the average performance of models across all benchmark images provides an overall ranking over models, here we consider a finer-grained measure of performance, first at the level of image categories, and then at the level of individual images. Analyzing the score breakdown of DeepFix and BMS across the 20 distinct image categories of CAT2000, we find that DeepFix consistently outperforms BMS on all image categories. Figs. 2 and 12 include DeepFix and BMS performances according to Area under ROC Curve (AUC), Normalized Scanpath Saliency (NSS), and Information Gain (IG) metrics.

An even finer-grained analysis can be achieved by considering performance on a per-image basis. In Fig. 3 we plot the number of images in the MIT300 dataset for which different models are the top performers (offer the best saliency

[3] Code for evaluation was used from https://github.com/cvzoya/saliency

predictions on an individual image level). The DeepFix and SALICON models provide the best predictions across most images, and these models are used in the rest of the paper.

Note that of the top neural networks evaluated on the MIT300 dataset, DeepFix is the only model that has been submitted to, and evaluated on, the CAT2000 dataset (as of March 2016). Thus, DeepFix is the focus in the main paper. In this supplement, we provide some additional results on SALICON wherever performance scores are available.

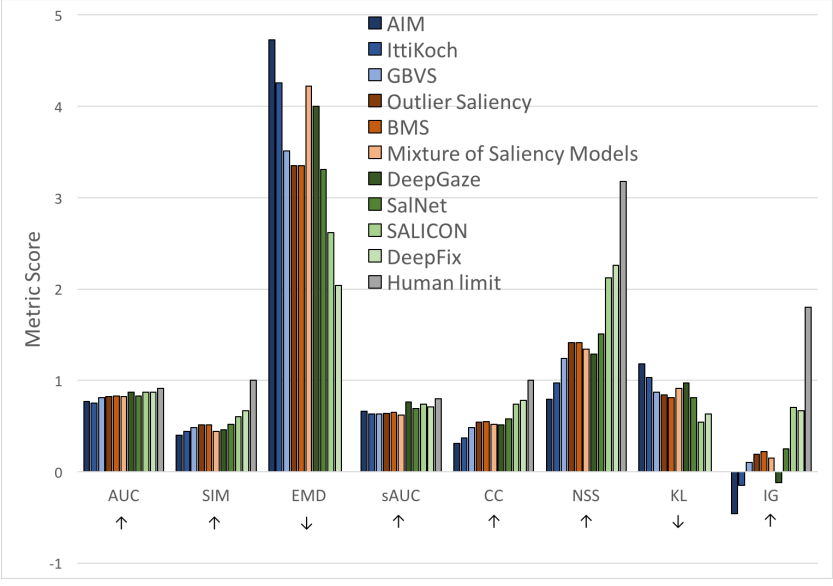


Fig. 1. In blue are three traditional bottom-up approaches (AIM [4], IttiKoch [5], GBVS [6]) commonly used for saliency comparisons, followed by in orange the top three non neural network approaches from the MIT300 benchmark (Outlier Saliency [7], BMS [8], Mixture of Saliency Models [9]), and in green the top four neural networks (DeepGaze [10], SalNet [11], SALICON [12], DeepFix [13]). The latter two models, SALICON and DeepFix, are the top-performing models across most metrics and are used for the main analyses in the paper. For all metrics except Earth Mover's Distance (EMD) and Kullback-Leibler divergence (KL), a higher score is better. The original, unscaled metric scores are included in each case. The human upper bound is measured as performance in the limit of infinite observers [3].

Saliency model	AUC ↑	sAUC ↑	NSS ↑	CC ↑	KL ↓	EMD ↓	SIM↑	IG ↑
DeepFix [13]	**0.87**	0.57	**2.29**	**0.88**	**0.41**	**1.11**	**0.75**	**0.20**
BMS [8]	0.85	**0.59**	1.67	0.67	0.83	1.95	0.61	-0.43

Table 1. Top-performing neural network and non-neural network models on CAT2000 Benchmark. Top scores are bolded. Lower scores for KL and EMD are better.

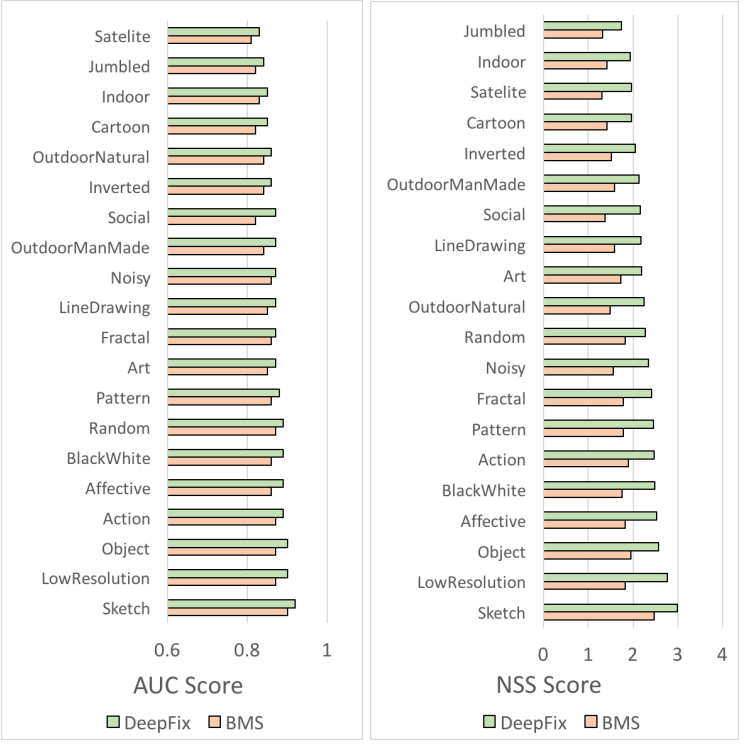


Fig. 2. The top neural network model (DeepFix) performs better across all 20 categories of CAT2000 than the top non-neural network model (BMS). The performance gain is especially noticeable with the Normalized Scanpath Saliency (NSS) metric, that takes into account false alarms as well as true positives. NSS offers a finer-grained comparison between models.

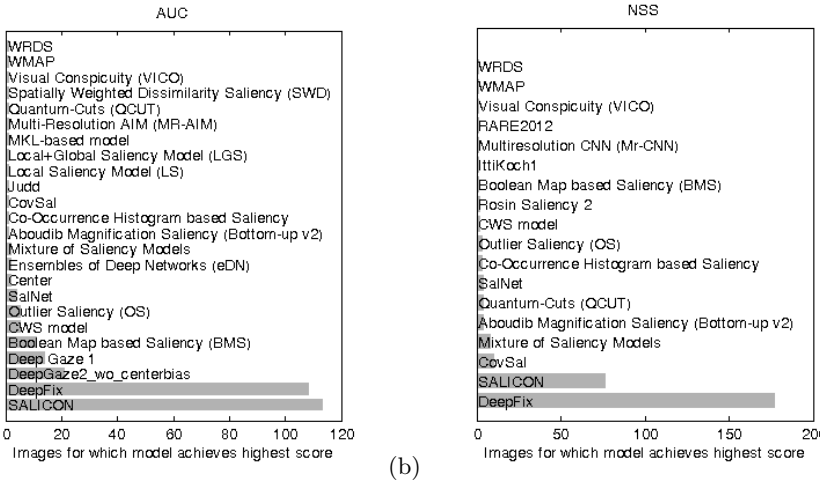


Fig. 3. A finer-grained analysis across all 300 benchmark images reveals that a single model does not always outperform all others. Here we see for how many of the 300 images each model achieves the highest scores for (a) Area under ROC Curve (AUC), (b) Normalized Scanpath Saliency (NSS).

Correlation-based greedy image selection

We greedily select one image at a time from the MIT300 dataset to best approximate the model score ranking on the MIT Saliency Benchmark, while keeping model performances on the images selected as uncorrelated as possible (to increase diversity of the subset of images selected). We employ correlation-based feature selection (CSF) [14], selecting a subset of k images by optimizing:

$$\begin{aligned} \text{CSF} &= \frac{k\overline{r_{st}}}{\sqrt{k+(k-1)\overline{r_{ss}}}} \\ \text{where } \overline{r_{st}} &= \frac{1}{k}\sum_{i=1}^{k} corr(s_i, t) \\ \overline{r_{ss}} &= \frac{k(k-1)}{2}\sum_{i=1}^{k}\sum_{j\neq i} corr(s_i, s_j) \end{aligned} \tag{1}$$

where s_i is a vector of the scores for all models on image i, and t is a vector of the scores for all models averaged over all 300 images of the MIT benchmark. The subset of k images chosen is such that model scores on the subset of images approximates the model scores on the whole benchmark. At the same time, the images are chosen to be diverse in scoring models. The 10 representative images selected in the main paper were chosen by optimizing Pearson correlation on the NSS metric.

Quantifying where people and models look in images

Defining regions of interest

Given discrete human fixation data (locations of all observer fixations) on an image, continuous fixation maps were obtained by convolving each fixation with a Gaussian with sigma equal to one degree of visual angle, to approximate the size of the human fovea and account for measurement errors.

We thresholded fixation heatmaps at the 95th percentile and collected all the connected components: a total of 999 regions from 300 images. Each connected component represents a region in an image with a density value in the top 5% of the heatmap. We then further filtered these regions by the number of fixations falling within them. We kept only the regions that had more than 5% of the total fixations on an image, leaving a total of 651 regions. On average, this worked out to be 9-12 fixations per region. Note that a single observer made about 5-6 fixations per image. Thus, even if all of a single observer's fixations landed on a region, that would not be enough to select it. In such a way we filtered out outliers, regions fixated by only one observer, and artifacts introduced during the connected component analysis. Artifacts can be caused by high-density regions occurring at the intersection of multiple highly-fixated spots due to constructive interference (when binary maps of fixations are smoothed via Gaussian blurring into continuous fixation maps).

Annotation tasks

Given regions of interest, the next step was to label them. To avoid experimenter subjectivity, labels for all the regions were crowdsourced, and a majority vote policy used. In this section we give additional details of the Amazon Mechanical Turk (MTurk) tasks described in the main paper, as well as the post-processing used to prepare the data labels for analysis. For the first MTurk annotation task, we showed workers images with specific regions of interest highlighted (one per image) and asked them to select all the labels that apply in describing each region. A single MTurk HIT consisted of 30 images. The following labels were provided: *Face*, *Part of face*, *Head*, *Person*, *Part of a person*, *Crowd of people*, *Legible text*, *Ilegible text*, *Non-english text*, *Symbol*, *Animal face*, *Part of an animal*, *Object*, *Background*, *Other*. We gathered annotations from 20 distinct workers to obtain robust labeling of regions. Majority vote was used to assign labels to regions, and multiple labels were used in the case of ties. In Table 2 we include the raw counts of all the resulting region types by label.

Fig. 4 is a multidimensional scaling visualization of the 651 regions that were labeled. Each region was represented with a histogram of the labels it received from the MTurk tasks before being projected onto the two dimensions visualized. Each point represents a labeled region. Nearby points correspond to regions that received a similar (set of) label(s). We can see that regions corresponding to parts of people cluster together, regions corresponding to text cluster separately, and the rest of the labels (object, background, and other) form a third cluster. For

further analyses, related labels were aggregated to have sufficient instances per label type. We used the following rules to aggregate labels:

- *Face* counted as *Part of face*
- *Person*, *Head*, and *Crowd* counted as *Part of person*
- *Legible text*, *Illegible text*, *Non-english text*, and *Symbol* counted as *Text*
- *Animal face* counted as *Part of animal*
- *Object*, *Background*, and *Other* counted as *Other*

Region type	Number of instances
Object	264
Part of a person	97
Legible text	84
Part of a face	67
Part of an animal	42
Crowd of people	33
Face	27
Other	19
Person	14
Background	13
Animal face	6
Illegible text	5
Head	5
Non-english text	3
Symbol	2

Table 2. What do people look at in images? Regions in images receiving a high density of eye fixations were labeled by MTurk workers. We summarize the 681 labels assigned to the 651 regions by MTurk workers.

A second MTurk task was designed to gain a better understanding of the regions that could not be easily described by the labels provided in the first MTurk task (regions that were annotated as *Object*, *Background*, and *Other*). Workers were presented with yes/no questions about image regions. A given MTurk worker would answer a specific question for a set of 30 images.

1. [**Object of gaze:**] Are any of the people in the image looking at something inside the highlighted region?

2a. [**Object of action:**] Are any people in the image interacting with something inside the highlighted region?

2b. [**Object of action:**] Is there an object inside the highlighted region that people are using in some way?

3. [**Unusual element:**] Is there something unusual about what is inside the highlighted region?

4a. [**Part of main subject:**] Is the highlighted region part of the main subject of this photograph?

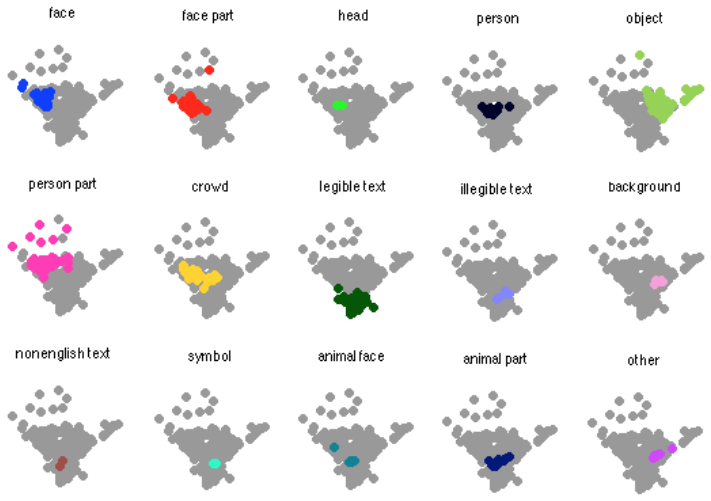


Fig. 4. Multidimensional scaling of the 651 labeled image regions (each region represented as a point), using the histogram of labels assigned to each region to project the region onto two axes. We color the points to mark regions that received a majority vote of each label type. We can see that regions corresponding to parts of people cluster together near the left corner, regions corresponding to text cluster separately near the bottom, and the rest of the labels (object, background, and other) form a third cluster in the right corner.

4b. [**Part of main subject:**] Is the highlighted region part of the photographer's main focus?
5a. [**Possible location for a person:**] If this was a video, could a person appear in the highlighted region of the image in the next instant?
5b. [**Possible location for a person:**] Would you expect to find a person in the highlighted region of an image (even if there's no one there now)?
6. [**Possible location of action/motion:**] If this was a video, could something move into the highlighted region of the image in the next instant?
7a. [**Location of action/motion:**] If this was a video, would there be an action or a motion happening inside the highlighted image?
7b. [**Location of action/motion:**] Is there an action happening inside the highlighted region?

A pilot task was run by asking every question about every image region, and collecting 3 MTurk responses per question. We kept the questions that generated worker agreement on at least 60% of the total regions. After this filtering round, the following questions were selected for final analysis: **1, 2b, 3, 4b, 5b, 7b**. For these selected questions, we collected additional responses from MTurk workers to have a total of 20 responses per question per region. Majority vote was used to add additional labels to image regions. For instance, if a majority of workers

replied affirmatively to question 1 when asked about a specific image region, then that image region would receive the label *Object of gaze*.

To combine the annotations from the first and second MTurk tasks, for any region that was labeled as *Object* in task 1, but also as *Object of action or gaze* in task 2, the original *Object* label was dropped, since the goal was to label each region as concretely as possible. Furthermore, *Other* labels were dropped in cases where a region received another, more concrete label. For any region labeled as *Part of a person or face*, the *Possible location for a person* label was dropped since the actual presence of a person made this latter labeling obsolete.

What do models miss?

Recall that regions of interest were selected from images because they were highly fixated by human observers: these regions occurred in the 95th percentile of human ground truth fixation maps. To determine if saliency models made correct predictions in these regions, we calculated whether the saliency in these regions was within the 95-th percentile of the saliency map for the whole image. Specifically, per model per region of interest, we computed the mean saliency value assigned by a model to each region. If the mean saliency value in the region of interest fell below the 95-th percentile of the saliency map, we counted this region as under-predicted by the model relative to the human ground truth. We then tallied up the types of regions that were most commonly under-predicted by models. In Fig. 5 we include a histogram of the failure modes of 10 different saliency models. From the MIT Saliency Benchmark, we selected 4 of the top neural network models, 3 of the top non neural network models, and 3 bottom-up models that are commonly used in saliency evaluations. We can see that the absolute number of failures of each types is significantly smaller among the SALICON and DeepFix models in comparison to all the rest, while the traditional bottom up models (AIM and IttiKoch make the poorest predictions).

Instead of counts, in Table 3 we compute the percent of the under-predicted regions that are due to each failure type (as in the main paper, but with more models). The percent is calculated in reference to the total number of patches, but since a patch can have multiple labels, these values do not add up to 100%. We include values for the top neural network models and top non neural network models. Although the absolute counts of errors are higher for the non neural network models, the makeup of failure modes is similar. Without an explicit face detector (as in the Mixture of Saliency Models), non neural network models miss many more faces in images than neural network models. Neural network models are also better at detecting text. In Table 4 we vary the percentile threshold from 75% to 95% to demonstrate that the distribution of failure modes remains relatively stable across thresholds. Using a percentile threshold instead of a fixed threshold accounts for model false alarms, so that a model is not at an advantage by spreading density across many regions in an image.

Because the label 'main' accounts for such a large percentage of the regions annotated across datasets, in Table 5 we provide a breakdown using other labels

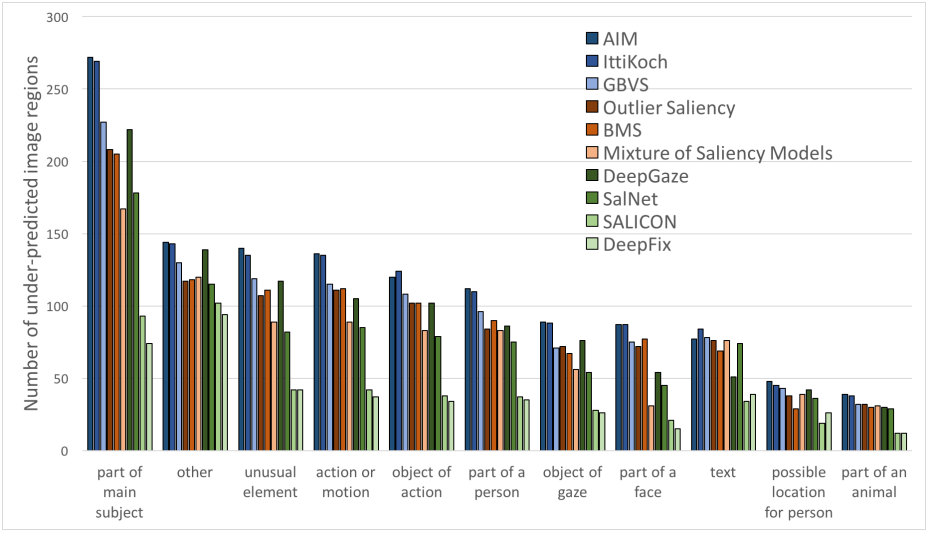


Fig. 5. Counts of model failures: image regions whose saliency is under-predicted relative to human ground truth fixation maps, aggregated by region type. The types of failures are similar across models, although the absolute quantity is significantly lower across neural network models (particularly the two models analyzed in this paper: SALICON and DeepFix).

Model	DeepFix [13]	SALICON [12]	DeepGaze [10]	SalNet [11]	BMS [8]	Mixture [9]	Outlier [7]
Part of main subject	31%	36%	48%	43%	46%	40%	46%
Unusual element	18%	16%	25%	20%	25%	22%	24%
Location of action/motion	16%	16%	23%	21%	25%	22%	24%
Text	16%	13%	11%	18%	16%	18%	17%
Part of a person	15%	14%	19%	18%	20%	20%	19%
Possible location for a person	15%	7%	9%	9%	7%	9%	8%
Object of action	14%	15%	22%	19%	23%	20%	22%
Object of gaze	11%	11%	16%	13%	15%	14%	16%
Part of a face	6%	8%	12%	11%	17%	8%	16%
Part of an animal	5%	5%	6%	7%	7%	8%	7%
Other	40%	40%	30%	28%	27%	29%	26%
Number regions	237	256	462	412	445	413	454

Table 3. Labels for under-predicted regions on MIT300 dataset. Regions are considered under-predicted if their predicted saliency falls below the 95-th percentile threshold. Percentages are computed over 681 labels assigned to 651 regions.

Model	DeepFix			SALICON		
Percentile threshold	**75**	**85**	**95**	**75**	**85**	**95**
Part of main subject	17%	32%	31%	33%	37%	36%
Unusual element	17%	16%	18%	19%	17%	16%
Location of action/motion	17%	18%	16%	23%	19%	16%
Text	13%	13%	16%	27%	24%	13%
Part of a person	17%	18%	15%	16%	16%	14%
Possible location for person	4%	7%	11%	11%	10%	7%
Object of action	8%	10%	14%	17%	18%	15%
Object of gaze	8%	12%	11%	11%	12%	11%
Part of a face	4%	3%	6%	9%	10%	8%
Part of an animal	8%	7%	5%	3%	5%	5%
Other	46%	41%	40%	28%	27%	40%

Table 4. Labels for under-predicted regions on MIT300 dataset by the top-performing models. Regions are considered under-predicted if their predicted saliency falls below the given percentile threshold. Here we show results on 3 different thresholds per model. The proportions of different types of mistakes models make remains relatively stable across different thresholds.

for all regions labeled 'main'. Recall that many regions have multiple labels assigned to them (thus percentages do not add up to 100%).

What can models gain?

As described in the main text, we modify saliency predictions by combining saliency maps with ground truth fixation maps in specific regions. Fig. 6 depicts this process. In Tables 6-7 we provide the results for the DeepFix and SALICON models on the MIT300 benchmark when specific image regions from the model predictions are replaced with ground truth. We provide the final metric scores obtained when the modified models are evaluated on the benchmark. Because different metrics have different ranges, we measure the percent gain in performance of each model with each modification relative to the human limit under that metric. In other words, we measure how much of the remaining gap to human ground truth each modification captures. These tables are meant to provide a general sense of the possible expected performance boost if different prediction errors are ameliorated.

Because of the significance of faces and text attracting significant observer fixations [15], we separately annotated bounding boxes around all faces and text in the MIT300 dataset. This provides a more complete set of instances for these regions than the MTurk-labeled pre-filtered image regions. In the last two rows of Tables 6-7 we provide the results for modified DeepFix and SALICON models, respectively, where we use annotated bounding boxes of faces to insert the ground-truth fixation maps for face regions into the saliency maps. We do the same with bounding boxes of text. This gives us an approximation of how well

Dataset	**MIT300**		**CAT2000**			
Model	**DeepFix**	**SALICON**	**DeepFix**			
Image category	**All**		**Social**	**Action**	**Indoor**	**Outdoor**
Unusual element	46%	39%	41%	75%	0%	33%
Location of action/motion	39%	33%	79%	86%	33%	11%
Text	9%	9%	3%	7%	0%	11%
Part of a person	26%	25%	18%	36%	17%	0%
Possible location for a person	16%	13%	9%	32%	17%	22%
Object of action	35%	31%	35%	64%	0%	11%
Object of gaze	27%	24%	71%	57%	0%	0%
Part of a face	15%	15%	56%	4%	0%	0%
Part of an animal	11%	8%	3%	7%	0%	0%
Other	14%	24%	0%	7%	67%	44%

Table 5. Understanding the breakdown of image regions labeled 'main'. The last row is the set of regions that were only assigned the 'main' label and can not be explained by anything else.

a model might be able to do on the saliency benchmark if it correctly predicted these special types of regions in images.

Note that for some metrics, performances do not always improve. Replacing certain regions of a saliency map with regions from the ground truth map can skew the overall distribution of saliency values and affect distribution-based metrics. For instance, the Earth Mover's Distance (EMD) metric highly penalizes extra density, and prefers sparser saliency maps even if the prediction does not exactly overlap with the ground truth [3]. Some form of histogram matching may be required during the composition procedure. This is a separate modeling consideration that is beyond the scope of this paper. The present analyses provide a rough approximation of the expected gains in performance if certain regions were correctly predicted. At the same time, the AUC metrics appear to have saturated and do not significantly change with model modifications. As can be seen from Fig. 1, even quite different saliency models have similar AUC scores. We believe this score is not providing a fine-grained comparison between saliency models, and a metric such as Normalized Scanpath Saliency (NSS) or Information Gain (IG) may be more appropriate for benchmarking model [3].

The importance of people

Some additional examples where the saliency of faces is either underestimated or overestimated are provided in Fig. 7.

Not all people in an image are equally important

Fig. 8 contains additional examples of cases where predicting the correct relative importance of people in an image is important for scene understanding. This

Fig. 6. Modified saliency maps were created by replacing the predicted saliency values within a region of interest with the ground-truth fixation map values at those locations. Separate analyses were conducted for different types of image regions. Visualized here (top to bottom) are regions containing a face, text, and object of gaze and action.

Saliency model	AUC ↑	sAUC ↑	NSS ↑	CC ↑	KL ↓	EMD ↓	SIM ↑	IG ↑
DeepFix (orig)	0.87	0.71	2.26	0.78	0.63	2.04	0.67	0.67
+ main subject	0.88	0.73	2.49	0.82	0.58	1.88	0.69	0.78
	+20.0%	+20.0%	+22.3%	+18.2%	+7.9%	+7.8%	+6.1%	+9.7%
+ unusual	0.87	0.72	2.39	0.80	0.60	1.95	0.68	0.73
	+0.0%	+10.0%	+12.6%	+9.1%	+4.8%	+4.4%	+3.0%	+5.3%
+ action/motion	0.87	0.72	2.37	0.80	0.60	1.96	0.68	0.72
	+0.0%	+10.0%	+10.7%	+9.1%	+4.8%	+3.9%	+3.0%	+4.4%
+ text	0.87	0.72	2.32	0.79	0.60	1.98	0.67	0.71
	+0.0%	+10.0%	+5.8%	+4.5%	+4.8%	+2.9%	+0.0%	+3.5%
+ person part	0.87	0.72	2.35	0.80	0.61	1.98	0.68	0.70
	+0.0%	+10.0%	+8.7%	+9.1%	+3.2%	+2.9%	+3.0%	+2.7%
+ location of person	0.87	0.72	2.30	0.79	0.62	2.00	0.67	0.69
	+0.0%	+10.0%	+3.9%	+4.5%	+1.6%	+2.0%	+0.0%	+1.8%
+ objects of action	0.87	0.72	2.38	0.80	0.61	1.96	0.68	0.72
	+0.0%	+10.0%	+11.7%	+9.1%	+3.2%	+3.9%	+3.0%	+4.4%
+ objects of gaze	0.87	0.72	2.34	0.79	0.61	1.98	0.67	0.70
	+0.0%	+10.0%	+7.8%	+4.5%	+3.2%	+-46.1%	+0.0%	+2.7%
+ face part	0.87	0.71	2.31	0.79	0.62	2.02	0.67	0.69
	+0.0%	+0.0%	+4.9%	+4.5%	+1.6%	+1.0%	+0.0%	+1.8%
+ animal part	0.87	0.71	2.30	0.79	0.62	2.01	0.67	0.68
	+0.0%	+0.0%	+3.9%	+4.5%	+1.6%	+1.5%	+0.0%	+0.9%
+ face boxes	0.87	0.72	2.34	0.79	0.61	2.08	0.67	0.69
	+0.0%	+10.0%	+7.8%	+4.5%	+3.2%	-2.0%	+0.0%	+1.8%
+ text boxes	0.87	0.72	2.34	0.80	0.60	1.98	0.67	0.72
	+0.0%	+10.0%	+7.8%	+9.1%	+4.8%	+2.9%	+0.0%	+4.4%
Human limit	0.92	0.81	3.29	1	0	0	1	1.80

Table 6. Improvements of DeepFix model on MIT300 if specific regions were accurately predicted. Performance numbers are over all 300 benchmark images, where regions from the ground truth fixation map are substituted into the DeepFix saliency maps (as in Fig. 6) to examine the change in performance. The percentage gain in performance is highlighted in gray, where the gain is computed as a fraction of the score difference between the original model score and the human limit.

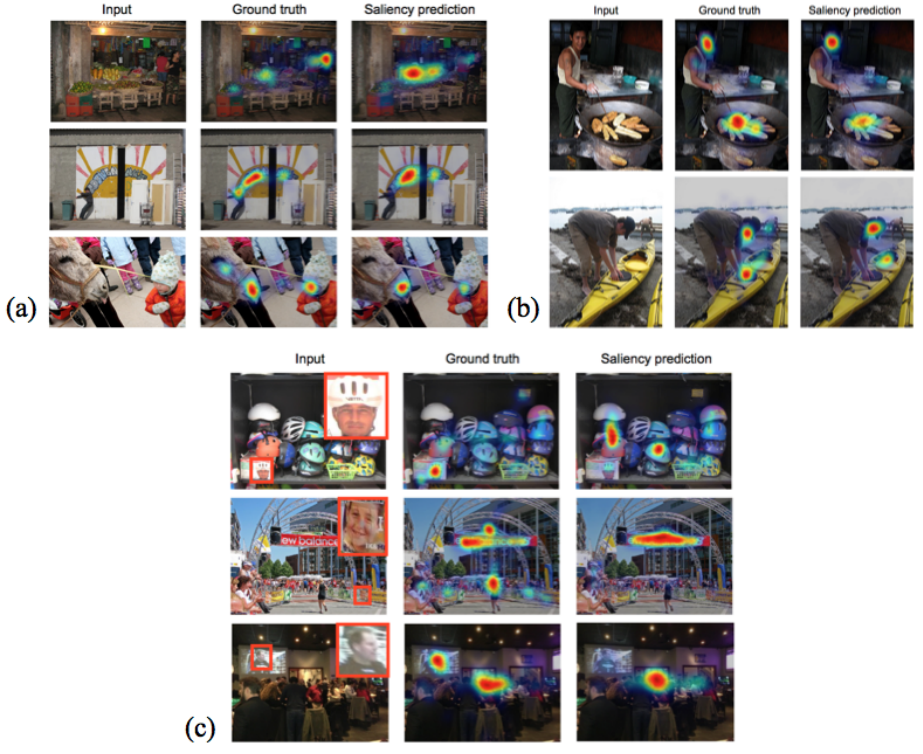


Fig. 7. Images from the MIT300 dataset containing a single visible face and the largest discrepancy between ground truth and model predictions. (a) Cases where the saliency model (DeepFix) underestimates the face saliency. These face images tend to be small, non-frontal, not centered in the image and otherwise harder to detect. (b) Cases where the saliency of the face in the image is over-estimated. In these images, people tend to fixate more the actions being performed than the individuals pictured. (c) Models are consistently failing to detect depictions of faces, in posters and photographs appearing within the input images. These faces often lack the context of a body, and appear at an unusual location in the image.

Saliency model	AUC ↑	sAUC ↑	NSS ↑	CC ↑	KL ↓	EMD ↓	SIM ↑	IG ↑
SALICON (orig)	0.87	0.74	2.12	0.74	0.54	2.62	0.60	0.71
+ main subject	0.87	0.75	2.27	0.77	0.53	2.61	0.59	0.75
	+0.0%	+14.3%	+12.8%	+11.5%	+1.9%	+0.4%	-2.5%	+3.7%
+ unusual	0.87	0.75	2.20	0.75	0.53	2.63	0.59	0.73
	+0.0%	+14.3%	+6.8%	+3.8%	+1.9%	-0.4%	-2.5%	+1.8%
+ action/motion	0.87	0.75	2.19	0.75	0.54	2.63	0.59	0.72
	+0.0%	+14.3%	+6.0%	+3.8%	+0.0%	-0.4%	-2.5%	+0.9%
+ text	0.87	0.74	2.15	0.75	0.53	2.61	0.59	0.71
	+0.0%	+0.0%	+2.6%	+3.8%	+1.9%	+0.4%	-2.5%	+0.0%
+ person part	0.87	0.74	2.17	0.75	0.53	2.63	0.59	0.72
	+0.0%	+0.0%	+4.3%	+3.8%	+1.9%	-0.4%	-2.5%	+0.9%
+ location of person	0.87	0.74	2.13	0.75	0.54	2.62	0.59	0.71
	+0.0%	+0.0%	+0.9%	+3.8%	+0.0%	+0.0%	-2.5%	+0.0%
+ objects of action	0.87	0.75	2.19	0.75	0.54	2.63	0.59	0.72
	+0.0%	+14.3%	+6.0%	+3.8%	+0.0%	-0.4%	-2.5%	+0.9%
+ objects of gaze	0.87	0.74	2.17	0.75	0.53	2.62	0.59	0.72
	+0.0%	+0.0%	+4.3%	+3.8%	+1.9%	+0.0%	-2.5%	+0.9%
+ face part	0.87	0.74	2.16	0.75	0.53	2.62	0.59	0.72
	+0.0%	+0.0%	+3.4%	+3.8%	+1.9%	+0.0%	-2.5%	+0.9%
+ animal part	0.87	0.74	2.14	0.75	0.53	2.62	0.59	0.71
	+0.0%	+0.0%	+1.7%	+3.8%	+1.9%	+0.0%	-2.5%	+0.0%
+ face boxes	0.87	0.74	2.19	0.75	0.55	2.72	0.59	0.70
	+0.0%	+0.0%	+6.0%	+3.8%	-1.9%	-3.8%	-2.5%	-0.9 %
+ text boxes	0.87	0.75	2.17	0.75	0.55	2.69	0.59	0.69
	+0.0%	+14.3%	+4.3%	+3.8%	-1.9%	-2.7%	-2.5%	-1.8 %
Human limit	0.92	0.81	3.29	1	0	0	1	1.80

Table 7. Improvements of SALICON model on MIT300 if specific regions were accurately predicted. Performance numbers are over all 300 benchmark images, where regions from the ground truth fixation map are substituted into the SALICON saliency maps to examine the change in performance. The first set of 10 region labels are obtained from the MTurk tasks. The next two rows (face and text boxes) were manually annotated to obtain a comprehensive set of each region.

figure demonstrates how model predictions deviate from human ground truth. The density of human fixations on a face is used as an approximation of the relative importance of that person in the image.

Objects of gaze and action

In the main text, we included some examples of objects of gaze that the best saliency models failed to predict. We demonstrated that a gaze-following model (explicitly trained to predict gaze [16]) can localize these regions in images. In Fig. 9, we include some additional examples of objects of gaze from the MIT300

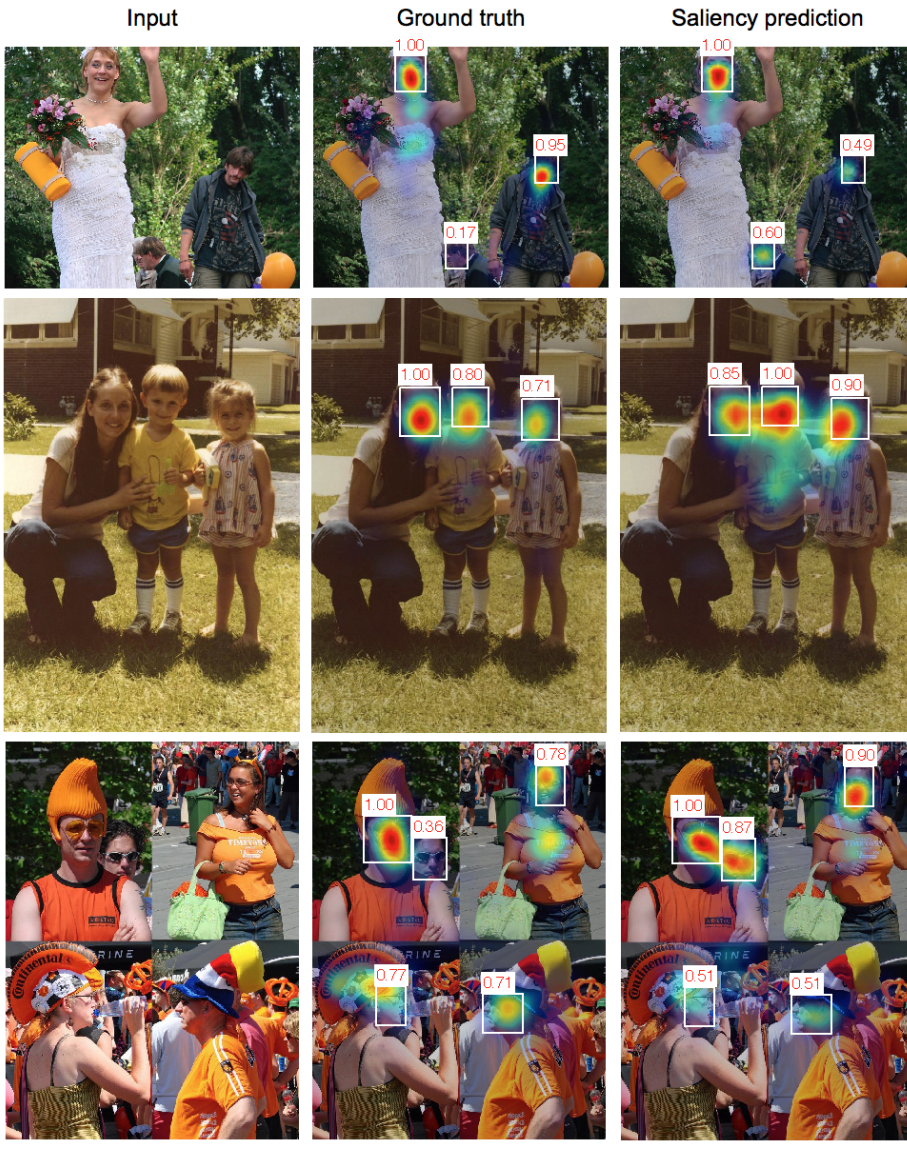


Fig. 8. Although recent saliency models have begun to detect faces in images with high precision, there is still a long way to go with regards to assigning the correct relative importance to different faces in an image. This requires an understanding of the interactions in an image: e.g., who is a likely participant in an action (top row) and who likely has the highest authority out of a group (second row). Facial expressions, accessories/embellishments, facial orientation, and position in a photo also contribute to the importance of individual faces (third row). In particular, image panels such as the last one can serve as good test cases for models.

dataset, but where at least one of the saliency models is able to localize them. We also provide the outputs from the gaze-following model on these images, to show its generalizability past the few examples provided in the main text.

The gaze saliency maps in the last column of Fig. 9 (and in the corresponding figure in the paper) were computed using the model and code provided by Recasens et al. [16]. This gaze-following model provides a prediction for the gaze of each of the subjects in the image. Its output consist of a heat map representing a combination among the different gaze predictions; that is, a map highlighting the objects people are looking at in the image. The procedure used to build the final gaze maps is described below.

1. Using face bounding boxes we compute the output of the model for each of the people in the image.
2. Using the importance score for each of the people in the image, we can weight each of the gaze maps by its relative importance in the image.
3. Adding up all the weighted maps we compute the final output map. These output maps provide a distribution over where each of the people is looking.

Note that the final weighted map captures the objects where people are looking and their relative importance to the full image.

In Fig. 10 we provide examples of objects of action for which gaze information can not be used to localize these regions in images, but body orientation and other body parts (specifically the hands) can help. This is a fruitful direction for further research.

Finer-grained datasets

Datasets broken up by image type, like CAT2000 on the MIT Saliency Benchmark, can help highlight images on which computational models might be able to achieve the greatest improvements. In Fig. 12 we include the performances the DeepFix model achieves on 20 image categories, measured as Information Gain [2] over the center baseline and over the best non neural network model (BMS). From this plot we can also see that DeepFix has made significant improvements over BMS across all image categories, with the biggest improvements across the non-natural images: sketches, patterns, and cartoons. However, deep models are not better predictors of human fixations, than a simple center baseline, on outdoor natural, pattern, and satellite images.

Another possible finer-grained test case for models is panel images: images constructed out of other images or image elements (Fig. 11). This can provide a direct measure for how models assign relative importance scores to different sub-images. It is a hard test case containing multiple salient objects, where the challenge is no longer to just detect the salient objects, but to determine which of the objects should have higher relative saliency. Models can be evaluated in this setting by comparing their saliency ranking of the image panels to a ground truth ranking (e.g. obtained by using maximum activations in fixation maps). Now that state-of-the-art saliency models have become very good detectors of

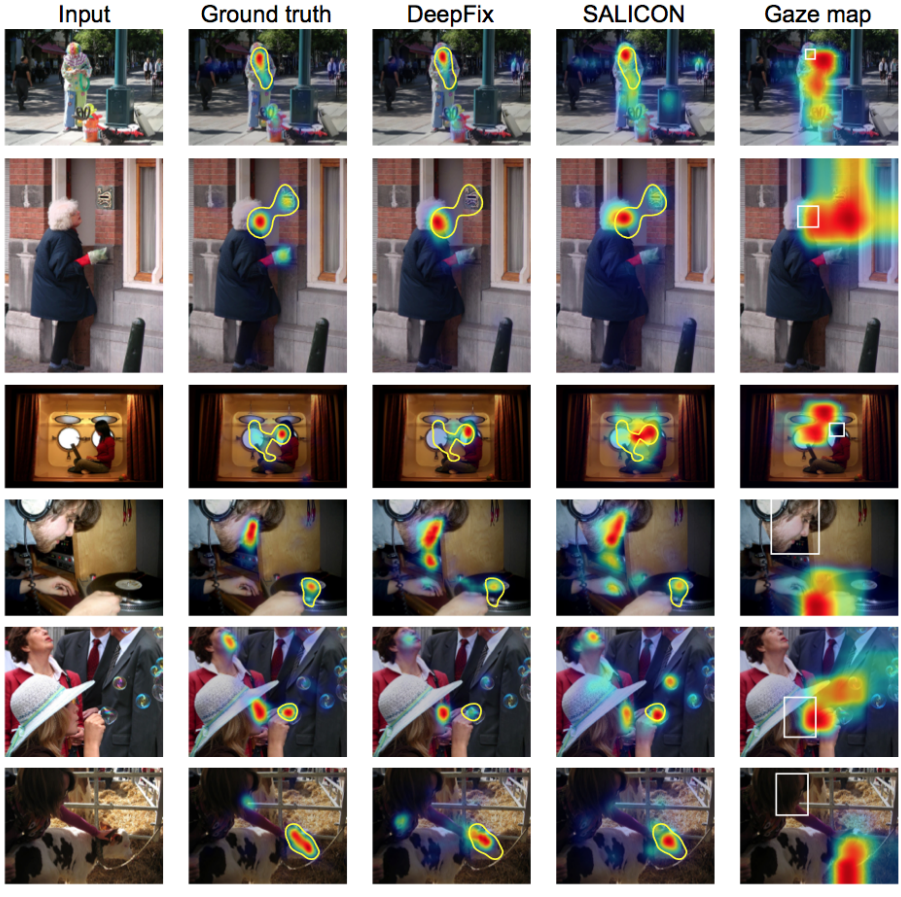


Fig. 9. Images in the MIT300 dataset labeled to contain objects of gaze but where at least one deep learning saliency model (DeepFix or SALICON) accurately predicts the human ground truth fixation map. The yellow outlines highlight high-density regions in the ground truth fixation map that were labeled by MTurk workers as regions on which the gaze of someone in the image falls. A model that explicitly predicts the gaze of individuals in an image can localize these objects of interest [16].

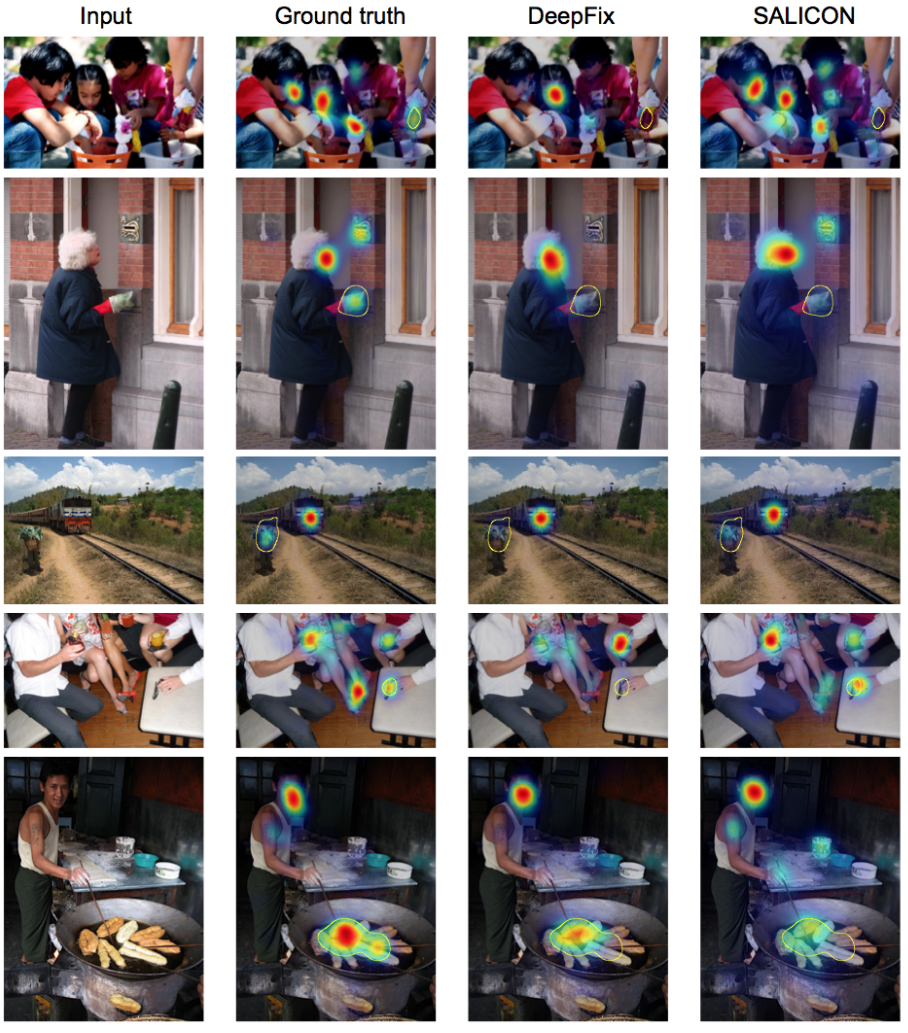
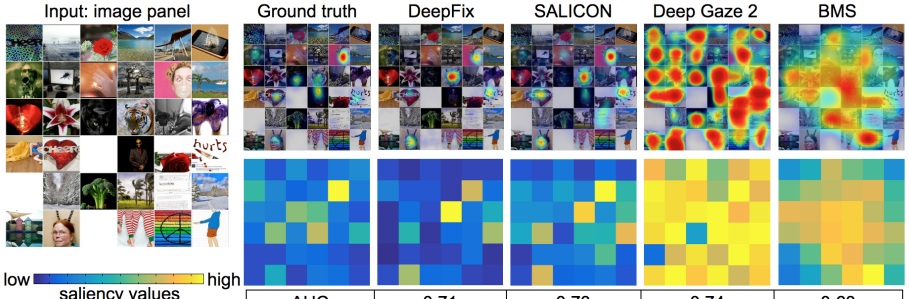


Fig. 10. Images in the MIT300 dataset labeled to contain objects of action - i.e. objects being acted on, or interacted with, by people in the scene. Included are images where both deep learning saliency models, DeepFix and SALICON, underestimate the saliency of these regions. Notice that the significance of these objects can not be inferred from gaze information (unlike in Fig. 9), since in all of these cases no one in the image is looking at the objects of interest. The yellow outlines highlight high-density regions in the ground truth fixation map that were labeled by MTurk workers as regions containing objects being acted on.

objects in images, the next step is to test them on their ability to predict relative importance of the objects, such as people in a crowd, a collection of text regions, or a set of panel images, which can call for new images, tasks, and evaluation procedures.



AUC	0.71	0.73	0.74	0.63
NSS	1.12	1.11	0.83	0.45
Panel ranking	0.62	0.68	0.39	0.33

Fig. 11. A finer-grained test for saliency models: determining the relative importance of different sub-images in a panel (this panel image is part of the MIT300 dataset). The top row contains the saliency map outputs given the input image panel. AUC and NSS scores are computed for these saliency maps. In the second row, the maximum response of each saliency model on each subimage is visualized (as an importance matrix). The importance ranking is a measure of the correlation of values in the ground truth and predicted importance matrices.

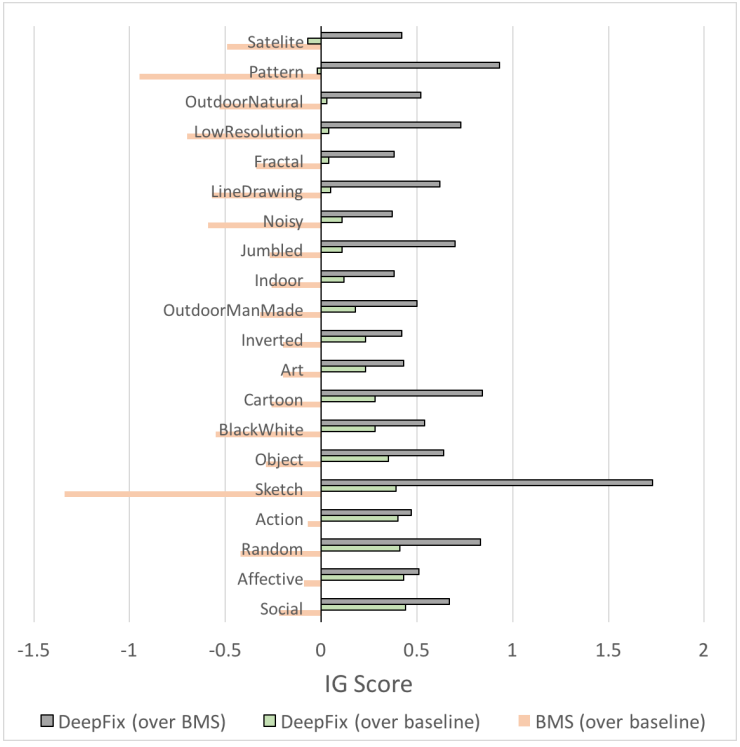


Fig. 12. Average Information Gain of DeepFix model over the center baseline and over the top-performing non neural network model (BMS), computed as in [2]. DeepFix consistently outperforms BMS across all image categories, with especially large gains on non-natural image categories like patterns, cartoons, and sketches. However, DeepFix still not provide prediction gains over the baseline for the first 3 categories, indicating its predictions might be missing key elements in those images.

Acknowledgments

This work has been partly funded by NSERC PGS-D Fellowship to Z.B., La Caixa Fellowship to A.R., NSF grant #1524817 to A.T., and Toyota Grant to F.D.

References

1. Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., Torralba, A.: MIT Saliency Benchmark. http://saliency.mit.edu/
2. Kümmerer, M., Wallis, T.S., Bethge, M.: Information-theoretic model comparison unifies saliency metrics. Proceedings of the National Academy of Sciences **112**(52) (2015) 16054–16059

3. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? arXiv preprint arXiv:1604.03605 (2016)
4. Bruce, N., Tsotsos, J.: Attention based on information maximization. Journal of Vision **7**(9) (2007) 950–950
5. Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Networks **19**(9) (2006) 1395 – 1407 Brain and Attention, Brain and Attention.
6. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems. (2006)
7. Chen, C., Tang, H., Lyu, Z., Liang, H., Shang, J., Serem, M.: Saliency modeling via outlier detection. Journal of Electronic Imaging **23**(5) (2014) 053023–053023
8. Zhang, J., Sclaroff, S.: Saliency detection: a boolean map approach. In: IEEE International Conference on Computer Vision. (2013)
9. Han, X., Satoh, S., Nakamura, D., Urabe, K.: Unifying computational models for visual attention yields better scores than state-of-the-art models. In: INCF Japan Node International Workshop: Advances in Neuroinformatics, AINI (2014)
10. Kümmerer, M., Theis, L., Bethge, M.: Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. arXiv preprint arXiv:1411.1045 (2014)
11. Pan, J., McGuinness, K., Sayrol, E., O'Connor, N., Giro-i Nieto, X.: Shallow and deep convolutional networks for saliency prediction. arXiv preprint arXiv:1603.00845 (2016)
12. Jiang, M., Huang, S., Duan, J., Zhao, Q.: Salicon: Saliency in context. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2015)
13. Kruthiventi, S.S., Ayush, K., Babu, R.V.: Deepfix: A fully convolutional neural network for predicting human eye fixations. arXiv preprint arXiv:1510.02927 (2015)
14. Hall, M.A.: Correlation-based feature selection for machine learning. PhD thesis, The University of Waikato (1999)
15. Cerf, M., Frady, E.P., Koch, C.: Faces and text attract gaze independent of the task: Experimental data and computer model. Journal of Vision **9**(12) (2009)
16. Recasens, A., Khosla, A., Vondrick, C., Torralba, A.: Where are they looking? In: Advances in Neural Information Processing Systems. (2015) 199–207