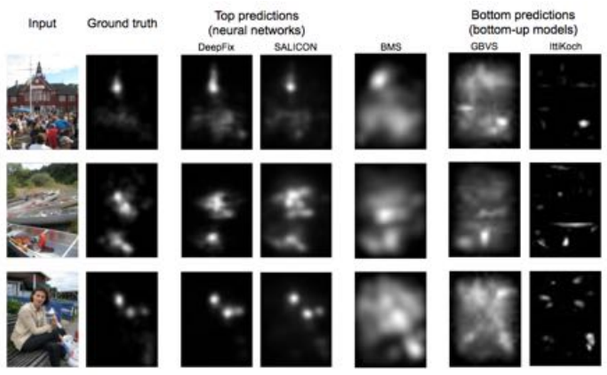
Where should saliency models look next?

Zoya Bylinskii, Adrià Recasens, Ali Borji, Aude Oliva, Antonio Torralba, Frédo Durand

NSERC CRSNG · Obra Social Fundación "la Caixa" · TOYOTA RESEARCH INSTITUTE · NSF

Have saliency models begun to converge on human performance? We re-examine the current state-of-the-art using a fine-grained analysis on image types, individual images, and image regions. We quantify up to 60% of remaining errors of saliency models. To continue to approach human-level performance, saliency models will need to discover higher-level concepts in images and reason about the relative importance of image regions.

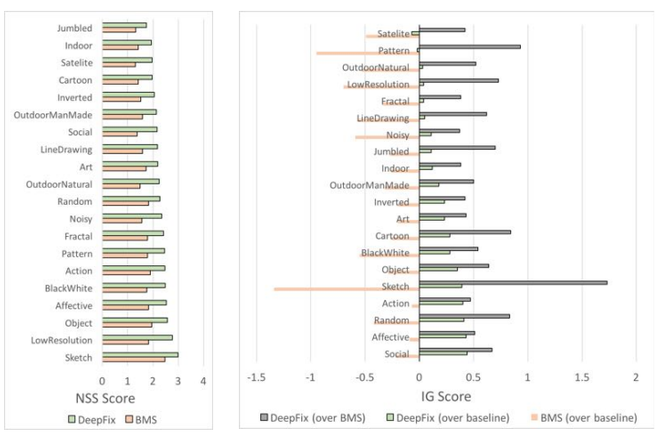
How far have **saliency models** come to ground truth?



Images most representative of model performance

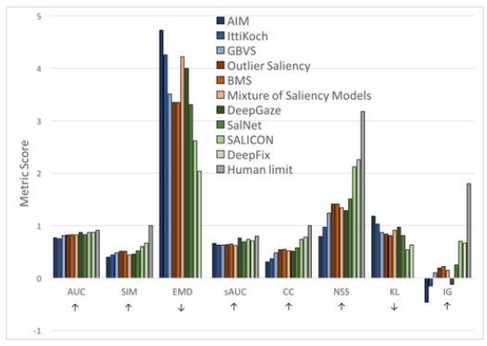
Saliency models have improved dramatically at ability to discover faces and text in images amidst clutter.

Finer-grained datasets and metrics



Finer-grained datasets can break up model performance by image category and uncover performance gaps.

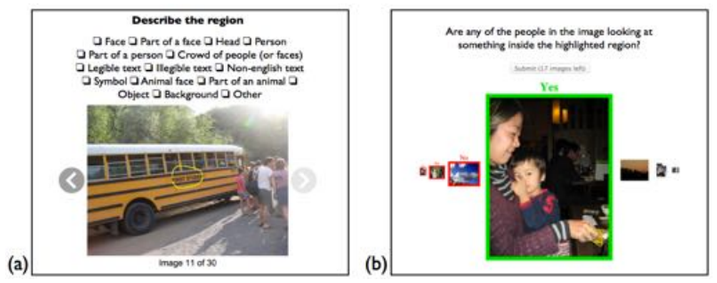
Monitoring performance on **MIT Saliency Benchmark**

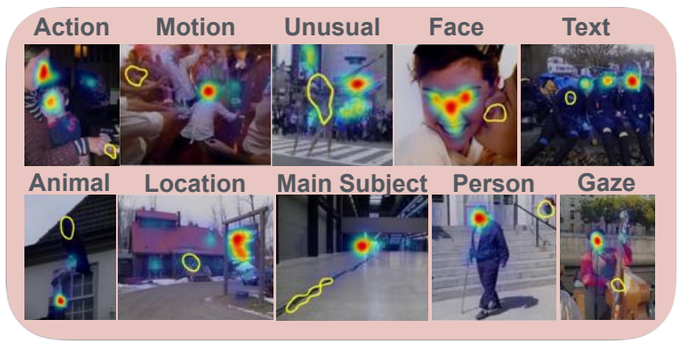


Bylinskii et al. What do different evaluation metrics tell us about saliency models? [arXiv 2016]
Kümmerer et al. Information-theoretic model comparison unifies saliency metrics [PNAS 2015]

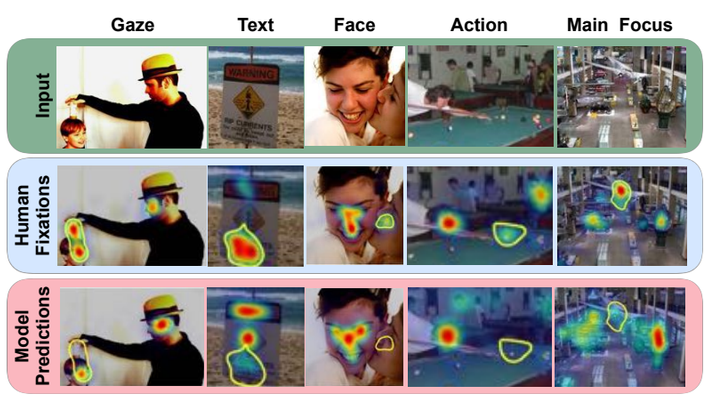
All state-of-the-art models are neural networks. Spikes in performances are observable on all metrics. Metrics like NSS and IG are more informative than others.

Crowdsourced annotations of highly-fixated image regions

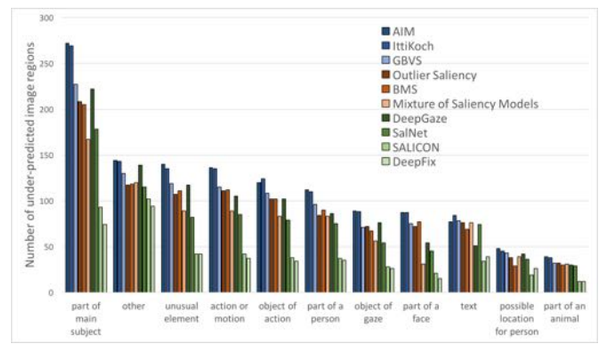




MTurkers labeled image regions corresponding to the 95th percentile of the human fixation maps (most fixated regions). Model predictions were overlapped with these regions to **quantify model errors.**



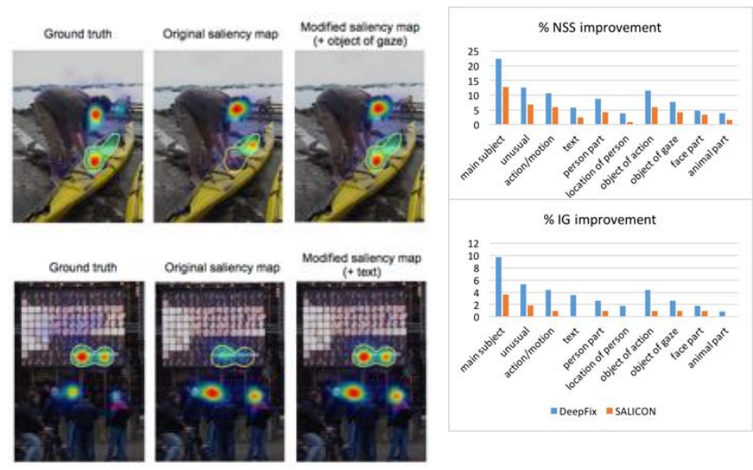
Aggregating **model errors**



Dataset	MIT300		CAT2000			
Model	DeepFix	SALICON	DeepFix			
Image category	All		Social	Action	Indoor	Outdoor
Part of main subject	31%	36%	49%	68%	12%	24%
Unusual element	18%	16%	33%	63%	8%	8%
Location of action/motion	16%	16%	67%	78%	8%	11%
Text	16%	13%	6%	5%	8%	29%
Part of a person	15%	14%	23%	37%	8%	5%
Possible location for a person	15%	7%	6%	24%	10%	11%
Object of action	14%	15%	27%	51%	0%	3%
Object of gaze	11%	11%	50%	44%	0%	0%
Part of a face	6%	8%	46%	7%	0%	0%
Part of an animal	5%	5%	3%	10%	0%	0%
Other	40%	40%	3%	2%	61%	37%

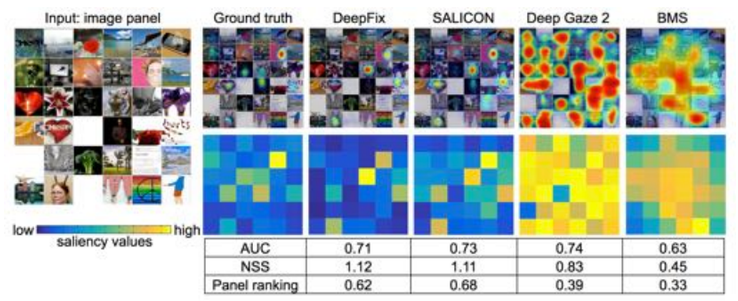
Types of errors made are common across models and datasets.

Imbuing predicted saliency maps with ground truth



Replacing saliency predictions in regions of interest with ground truth can **approximate performance gains** on MIT Saliency Benchmark.

Finer-grained tasks and evaluations



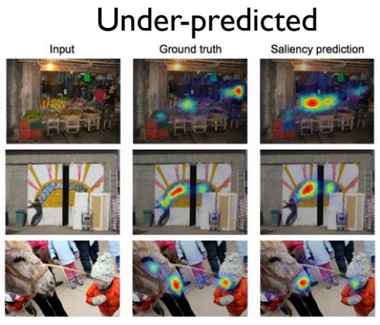
AUC	0.71	0.73	0.74	0.63
NSS	1.12	1.11	0.83	0.45
Panel ranking	0.62	0.68	0.39	0.33

Evaluating the **relative importance** of different image regions requires higher-level image understanding.

What are saliency models missing?

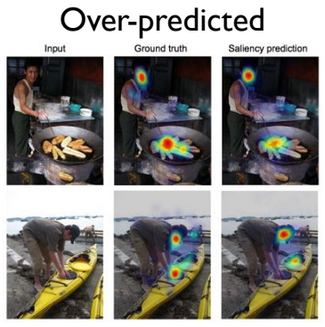
Assigning correct relative importance to **faces**

Under-predicted



Missed depictions

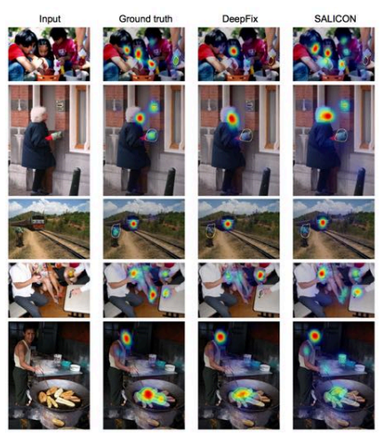
Over-predicted



Which **face** is most important?

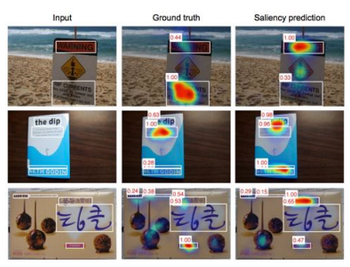
Current saliency models are good face detectors. The next challenge is analyzing the **relative importance** of faces compared to other faces and image content.

Objects of action: what is being acted upon?

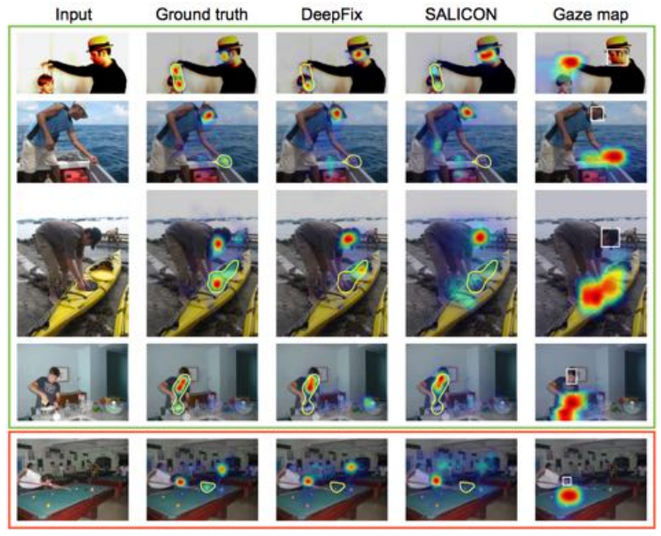


Which is the most important piece of **text**?

Which text in a scene provides the most relevant information for image understanding? At which point does saliency modeling become user-specific instead of populations-specific?



Objects of gaze: what is being looked at?



Recasens et al. Where are they looking? [NIPS 2015]

An explicit model of gaze can provide important cues not currently used by saliency models (above). In a similar manner, body posture and hand positions can point to objects of interest in a scene (left).