Consistent Saliency Benchmarking: How One Model Can Win on All Metrics

Matthias Kümmerer   Thomas S.A. Wallis   Matthias Bethge

Understanding how humans place their gaze is important for understanding how humans exlore their environment and for computer vision applications and has attracted research for many decades. So called "saliency models" compute a "saliency map" to predict fixations for an image. Many different saliency models have been proposed, from low-level feature integration to complex deep-learning based models, and more are added every year.

However the field is facing a fundamental problem: there is no agreed-upon metric for assessing the quality of a saliency map. Instead, e.g. the most commonly used MIT saliency benchmark evaluates a total of eight metrics which yield highly inconsistent model rankings. This has led to contradicting conclusions about which algorithms are most predictive. We have previously shown that treating models probabilistically and evaluating log-density saliency maps removes most of the disagreement, but at the price of performing suboptimally in most metrics.

Here we apply Bayesian utility theory to the problem: the metric is the utility function and the saliency map represents the particular choice one must make given the predicted fixation density and the utility function. We can show that it is impossible for one saliency map to perform well in all saliency metrics (Figure 1c). Instead we propose a principled way to derive different metric-dependent saliency maps from a model's predicted fixation density by maximizing the expected metric performance of a saliency map under the predicted fixation density (Figure 1b).

By converting several influential saliency map models to probabilistic models, we show that our approach indeed improves benchmarking on real data. We demonstrate that this allows models to compete with state-of-the-art for each metric (Figure 2). Furthermore, by using the right saliency maps for different metrics, the model ranking is consistent over different metrics. This indicates that the existing saliency metrics are in fact not measuring different properties of the predicted fixation densities – they just interpret saliency maps in different ways. This simple yet meaningful way to compare saliency models should allow a better understanding of which mechanisms work and which don't work and provide a better guide to future research.

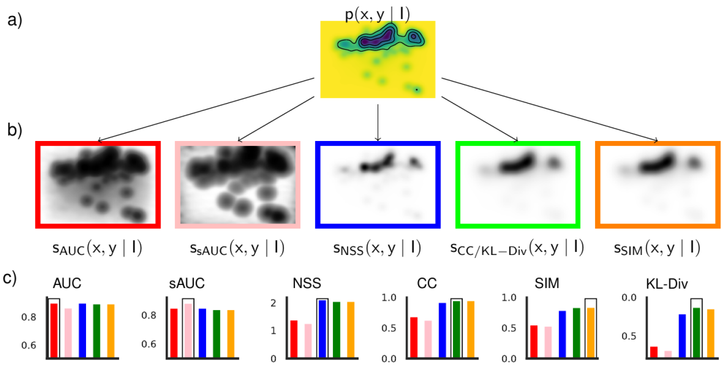

Figure 1: For a given fixation density (a) we show saliency maps optimal for different metrics (b). Every saliency map performs poorly in some of the metrics (c).

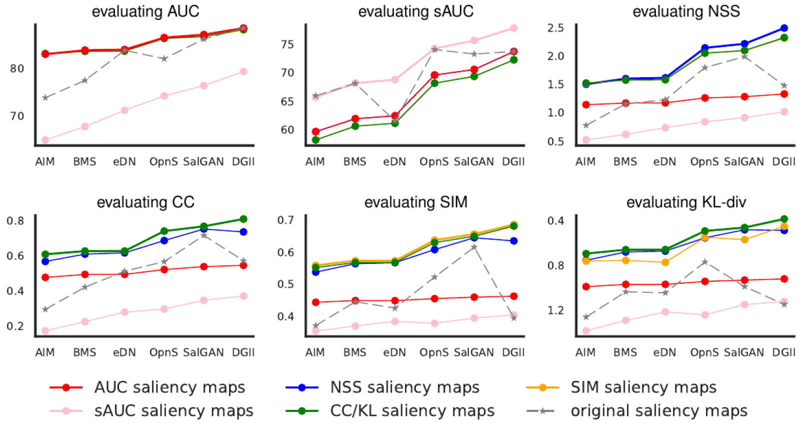

Figure 2: Evaluation on the MIT1003 dataset: The correct saliency map for each metric (thick lines) results in the best possible performance among all saliency map types (other lines) for each model and results in consistent model rankings across metrics – unlike the original saliency maps (dashed lines)