SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks

Xun Huang[1,2], Chengyao Shen[1], Xavier Boix[1,3], and Qi Zhao[1]

[1]Department of Electrical and Computer Engineering, National University of Singapore

[2]School of Computer Science and Engineering, Beihang University

[3]CBMM, Massachusetts Institute of Technology

Introduction

Saliency in Context (SALICON) is an ongoing effort that aims at understanding and predicting visual attention.

Motivation

Semantic Gap: Conventional saliency models typically rely on low-level image statistics to predict human fixations. While these models perform significantly better than chance, there is still a large gap between model prediction and human behavior. This gap is largely due to the limited capability of models in predicting eye fixations with strong semantic content, the so-called semantic gap.

Contributions

This poster presents a focused study to narrow the semantic gap with an architecture based on Deep Neural Network (DNN). Two key components of our architecture are:

- **Saliency Objective:** Objective functions based on the saliency evaluation metrics
- **Extension to Multi-scale:** Feature integration at different image scales

Compared with 14 saliency models on 6 public eye tracking benchmark datasets, results demonstrate that our DNNs can automatically learn features for saliency prediction that surpass by a big margin the state-of-the-art. In addition, our model **ranks top** under all seven metrics on the MIT300 challenge set.

Methods

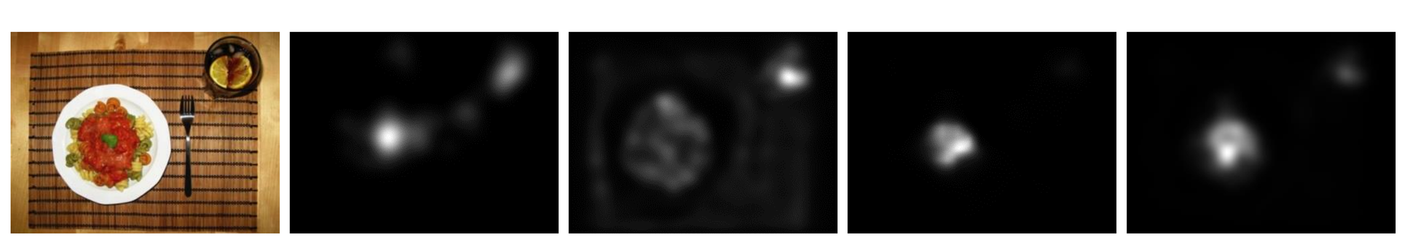
Saliency Objective

Four saliency evaluation metrics as objective functions of the back-propagation:

- Kullback-Leibler divergence (KLD)
- Normalized Scanpath Saliency (NSS)
- Similarity (Sim)
- Linear Correlation Coefficient (CC)

Extension to Multi-scale

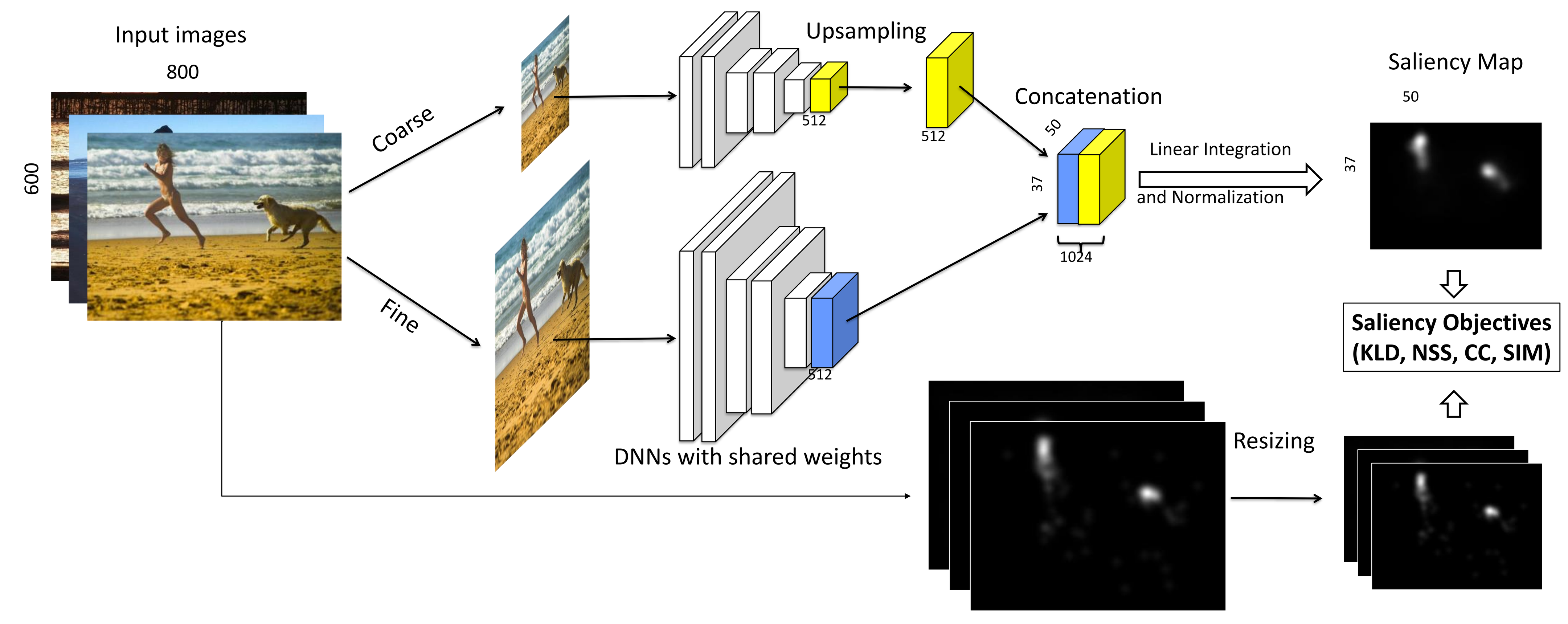
We use the input image at different scales obtained by downsampling the image. Each scale is processed by a DNN. The neural responses of all DNNs are concatenated to predict the saliency map. The effect of multi-scale is qualitatively illustrated below:



Stimuli | Human | Fine | Coarse | Fine+Coarse

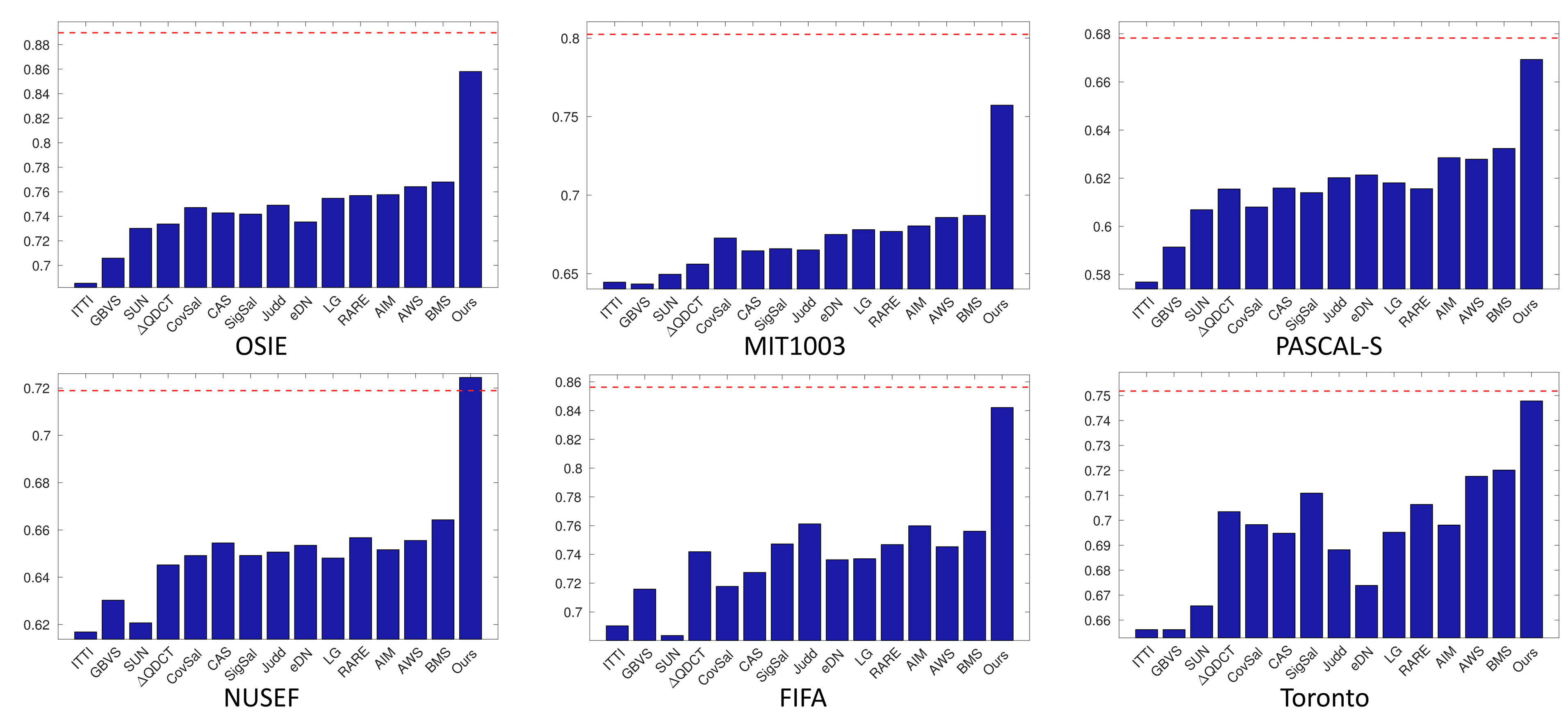
Methods

Architecture of the DNN

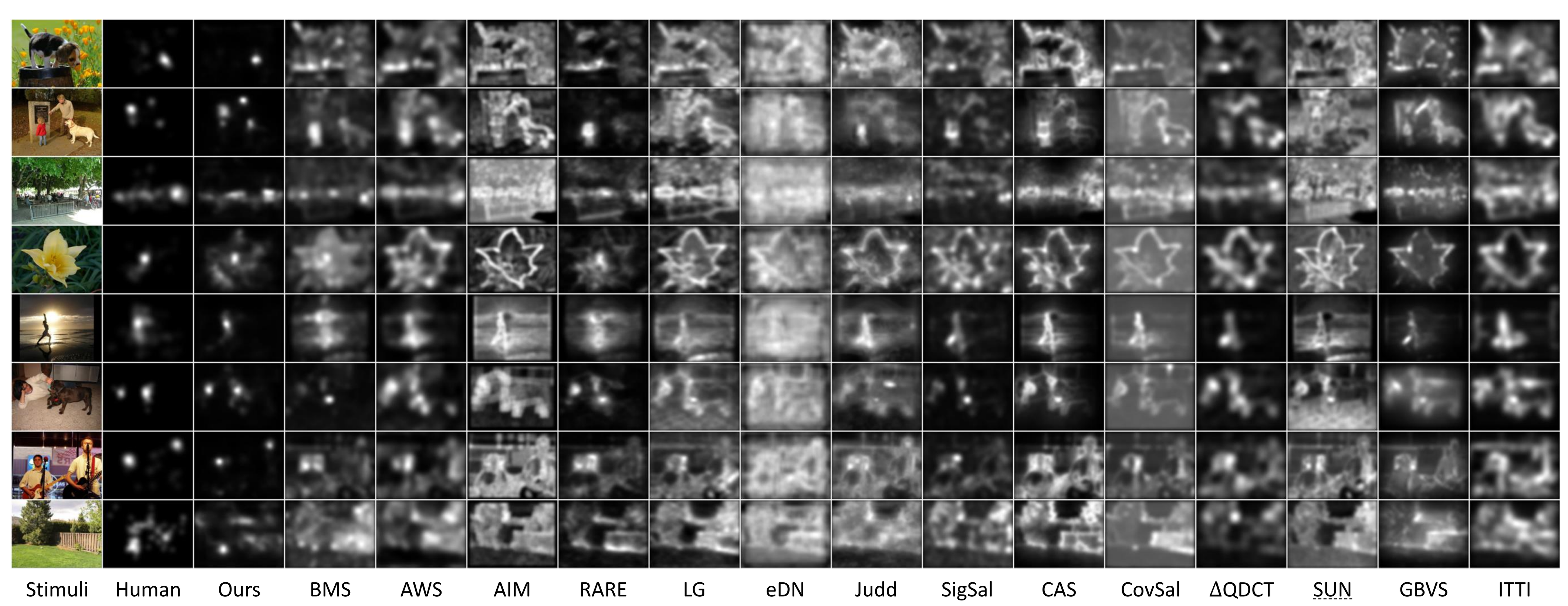


Results

Quantitative comparison (in sAUC) to the state-of-the-art models in 6 datasets.

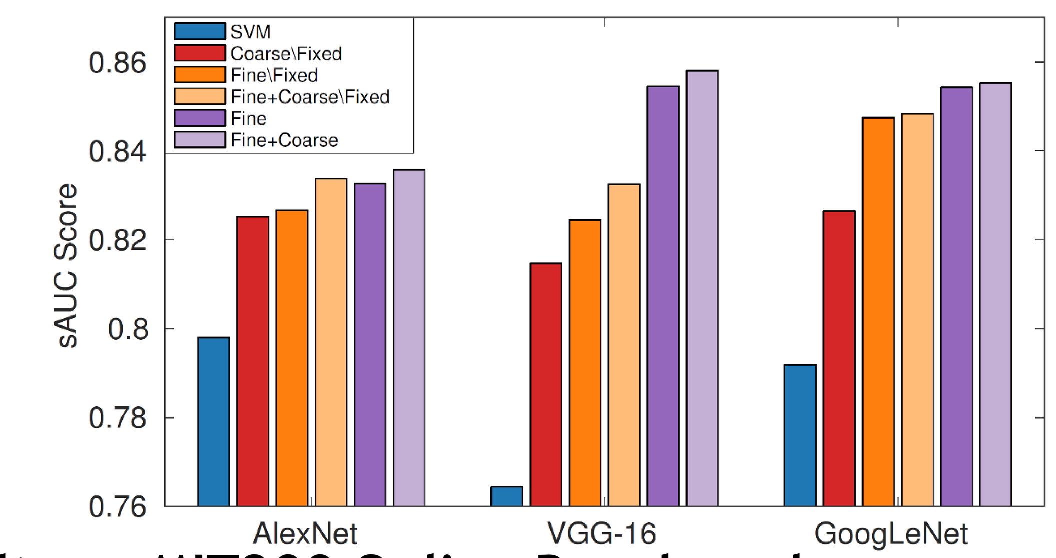


Qualitative comparison to the state-of-the-art



Results

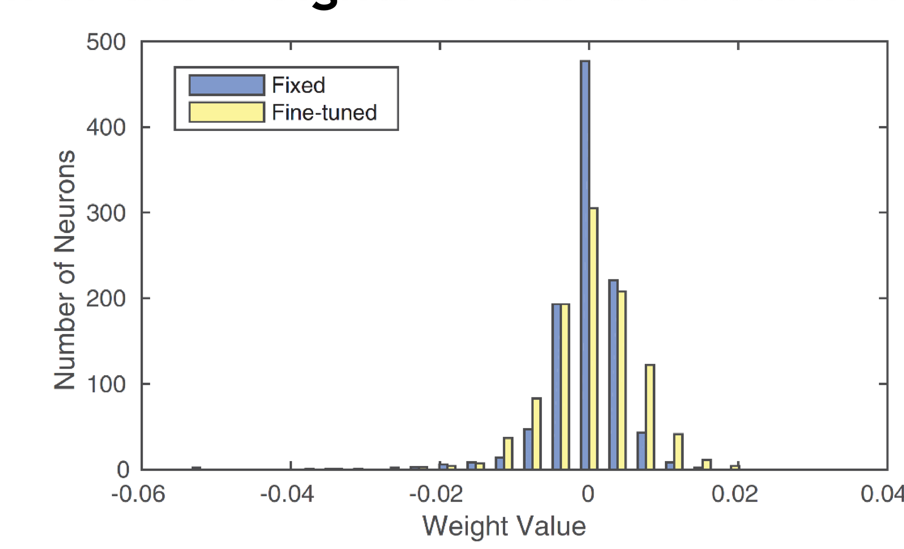
Intra-model comparison on multi-scale and fine-tuning



Results on MIT300 Online Benchmark (in 7 evaluation metrics)

Evaluation Metric	Gauss	State-of-the-art	Ours	Infinite Humans	Relative Advances
AUC-Judd	0.78	0.84 Deep Gaze	0.87	0.91	42.9%
AUC-Borji	0.77	0.83 Deep Gaze	0.85	0.87	50.0%
sAUC	0.51	0.68 AWS	0.74	0.80	50.0%
NSS	0.92	1.41 BMS	2.12	3.18	40.1%
CC	0.38	0.55 BMS	0.74	1	42.2%
Similarity	0.39	0.51 BMS	0.6	1	18.4%
EMD	4.81	3.33OS	2.62	0	21.3%

Histogram of the weights of the convolutional filter



Visualization

Top 3 salient features | Top 3 non-salient features

Scan the QR code to try our demo!