Unified Image and Video Saliency Modeling (Supplementary Material)

Richard Droste*, Jianbo Jiao*, and J. Alison Noble

University of Oxford
{richard.droste, jianbo.jiao, alison.noble}@eng.ox.ac.uk

1 Introduction

In this supplementary material, we provide additional quantitative and qualitative results for a better understanding of the proposed model for unified image and video saliency analysis. The contents are structured as follows:

2 Additional Qualitative Video Saliency Results

We present further qualitative video saliency prediction results in addition to those shown in the main paper. Also, we include comparisons to predictions generated with state-of-the-art methods [6,4]. Representative clips are sampled from the three video saliency datasets (DHF1K [6], UCF Sports [5], and Hollywood-2 [5]). The results are shown in the enclosed video file *video_3601.mp4* (also available at https://www.youtube.com/watch?v=4CqMPDI6BqE). Video frame-based examples are shown in Figure 1.

3 Additional Qualitative Images Saliency Results

We include further qualitative image saliency prediction results in addition to those presented in the main paper. Representative images are sampled from the SALICON [1] and MIT1003 [2] datasets. The results are shown in Figure 2 and Figure 3 for SALICON and MIT1003, respectively.

* Richard Droste and Jianbo Jiao contributed equally to this work.

Table 1. Ablation study of the domain-adaptive modules on the DHF1K and SALICON validation sets. The proposed components are added individually to a new baseline (*Baseline+...+Smoothing*) to quantify their contribution. Training setting (vi) is used for this study.

Config. \ Dataset	DHF1K					SALICON				
	KLD ↓	AUC-J ↑	SIM ↑	CC ↑	NSS ↑	KLD ↓	AUC-J ↑	SIM ↑	CC ↑	NSS ↑
Baseline + ... + Smoothing*	1.770	0.882	0.295	0.416	2.305	0.369	0.848	0.690	0.799	1.654
* + DABN	1.852	0.880	0.317	0.396	2.212	0.355	0.851	0.717	0.807	1.747
* + DA-Gaussians	1.748	0.884	0.315	0.412	2.278	0.386	0.848	0.679	0.794	1.647
* + DA-Fusion	1.706	0.888	0.326	0.434	2.437	0.326	0.854	0.712	0.820	1.750
* + DA-Smoothing	1.754	0.883	0.304	0.418	2.302	0.379	0.847	0.683	0.793	1.677
* + BypassRNN	1.784	0.882	0.322	0.412	2.302	0.356	0.853	0.695	0.819	1.721

4 Cross-Domain Predictions

Here, we analyze the impact of the domain-adaptive modules when predicting visual saliency on the same input. Results for video saliency prediction are shown in the second part of the attached video file *video_3601.mp4*. Figure 4 and Figure 5 show the results for image saliency prediction on SALICON and MIT1003 data, respectively. It is visible in Figure 4 that the video-specific settings (DHF1K, Hollywood-2, UCF Sports) cause the model to focus less on text and to focus on a single central object compared to the SALICON-specific setting. Similar observations can be made for the results shown in Figure 5.

5 Additional Center Bias Analysis

Here, we aim to evaluate the ability of the domain-adaptive learned Gaussian prior maps to capture the dataset-specific center biases. The results are shown in Figure 6. The upper row shows the averaged saliency targets for each training dataset as an approximation of the true center biases. In order to reveal the learned center biases, saliency predictions based on an all-zero input are generated for each set of domain-adaptive modules. For the video saliency datasets, the learned bias reflects the true biases visibly well. For SALICON, the true bias is significantly wider than the learned bias. A possible explanation is that the spread-out true bias for SALICON is not caused by a more spread-out center bias of the viewers, but rather by a spread-out placement of salient objects.

6 Additional Ablation Studies

In the main paper, we perform an ablation study on the components of the proposed methods. Here, we further ablate the individual domain-adaptive modules in Table 1. We use the same evaluation metrics as in the main paper and perform the study on the DHF1K and SALICON datasets. As a baseline for this study we use the *Baseline* model of the main ablation study with modules

Table 2. Cross-dataset generalization analysis on the SALICON benchmark test set. The training settings (i) to (vi) denote training with: (i) DHF1K, (ii) Hollywood-2, (iii) UCF Sports, (iv) SALICON, (v) DHF1K+Hollywood-2+UCF Sports, and (vi) DHF1K+Hollywood-2+UCF Sports+SALICON.

	KLD↓	AUC-J ↑	SIM ↑	s-AUC ↑	CC ↑	NSS ↑	IG ↑
Training setting (i)	0.45	0.83	0.65	0.67	0.75	1.61	0.43
Training setting (ii)	0.50	0.83	0.63	0.67	0.73	1.52	0.35
Training setting (iii)	0.82	0.81	0.61	0.67	0.65	1.42	0.00
Training setting (iv)	0.42	0.86	0.78	0.74	0.88	1.95	0.72
Training setting (v)	0.48	0.83	0.66	0.66	0.74	1.61	0.44
Training setting (vi)	0.35	0.86	0.78	0.74	0.88	1.95	0.78

added up to and including the *Smoothing* module. Then we add the individual domain-adaptive modules to this new baseline to analyze their respective effectiveness. Specifically, we add the domain-adaptive batch normalization (*DABN*), Gaussians (*DA-Gaussians*), Fusion (*DA-Fusion*), Smoothing (*DA-Smoothing*), and the Bypass RNN (*BypassRNN*). The results in Table 1 show that each domain-adaptive module contributes differently to the performance, in which the DA-Fusion contributes the most for both dynamic and static scenes. This is consistent with our analyses in the main paper which indicate that this module has an important contribution towards mitigating the domain shift.

7 SALICON Cross-Dataset Generalization

Here we analyze the cross-dataset generalization of the proposed UNISAL model for image saliency prediction on the SALICON benchmark test set. Specifically, we analyze the performance of our UNISAL model on the SALICON dataset when training with different datasets, *i.e.*, the six training settings described in the main paper, where setting (vi) is our final model. In this study, we follow the standard SALICON benchmark evaluation pipeline and include two additional metrics of KL-divergence (*KLD*) and Information Gain (*IG*). The results are shown in Table 2. We observe that the model performs slightly worse when training on video datasets only compared to training on SALICON, even when jointly training with the three video datasets (setting (v)). This observation confirms the existence of a domain shift between image and video saliency data. On the other hand, when jointly training with video and image datasets, the performance is boosted on some metrics while remaining stable on the others. This further validates the effectiveness of the proposed UNISAL approach to unify video and image saliency modeling.

8 Details for Quantitative Evaluation

8.1 Scoring SalEMA with Training Setting (vi)

For fairness of comparison, we score the SalEMA model[3] after fine-tuning it with training setting (vi), *i.e.*, DHF1K+Hollywood-2+UCF Sports+SALICON. For this, we use the official implementation provided by the authors under https://github.com/Linardos/SalEMA/. We fine-tune the *SalEMA30.pt* weights with the default training settings. SALICON images are treated as single-frame videos. The scores are computed on the test sets of UCF Sports and Hollywood-2 and the validation sets of DHF1K and SALICON, whose test sets are held-out for benchmarking.

8.2 Scoring ACLNet on SALICON

To obtain an additional baseline for image saliency prediction performance of an existing video saliency model besides SalEMA, we score the ACLNet model on the SALICON validation set (the test set is held-out for benchmarking). We compute the scores when using either the auxiliary image saliency prediction output or the LSTM output of the model. We find that the scores of the auxiliary output are better for all metrics and consequently report these in the paper.

8.3 Sources of Other Benchmark Scores

The scores of previous video saliency models on the DHF1K, UCF-Sports and Hollywood-2 datasets are obtained from [6]. The scores of the previous image saliency models on the SALICON and MIT300 benchmarks were obtained from the respective papers.

9 Code

The full code for evaluating and training the UNISAL model is available at https://github.com/rdroste/unisal.

References

1. Jiang, M., Huang, S., Duan, J., Zhao, Q.: Salicon: Saliency in context. In: CVPR (2015)
2. Judd, T., Durand, F., Torralba, A.: A Benchmark of Computational Models of Saliency to Predict Human Fixations. Mit-Csail-Tr-2012 **1**, 1–7 (2012)
3. Linardos, P., Mohedano, E., Nieto, J.J., McGuinness, K., Giro-i Nieto, X., O'Connor, N.E.: Simple vs complex temporal recurrences for video saliency prediction. In: BMVC (2019)
4. Pan, J., Ferrer, C.C., McGuinness, K., O'Connor, N.E., Torres, J., Sayrol, E., Giro-i Nieto, X.: Salgan: Visual saliency prediction with generative adversarial networks. arXiv:1701.01081 (2017)

5. Stefan Mathe, C.S.: Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. IEEE TPAMI **37** (2015)
6. Wang, W., Shen, J., Xie, J., Cheng, M.M., Ling, H., Borji, A.: Revisiting video saliency prediction in the deep learning era. IEEE TPAMI (2019)

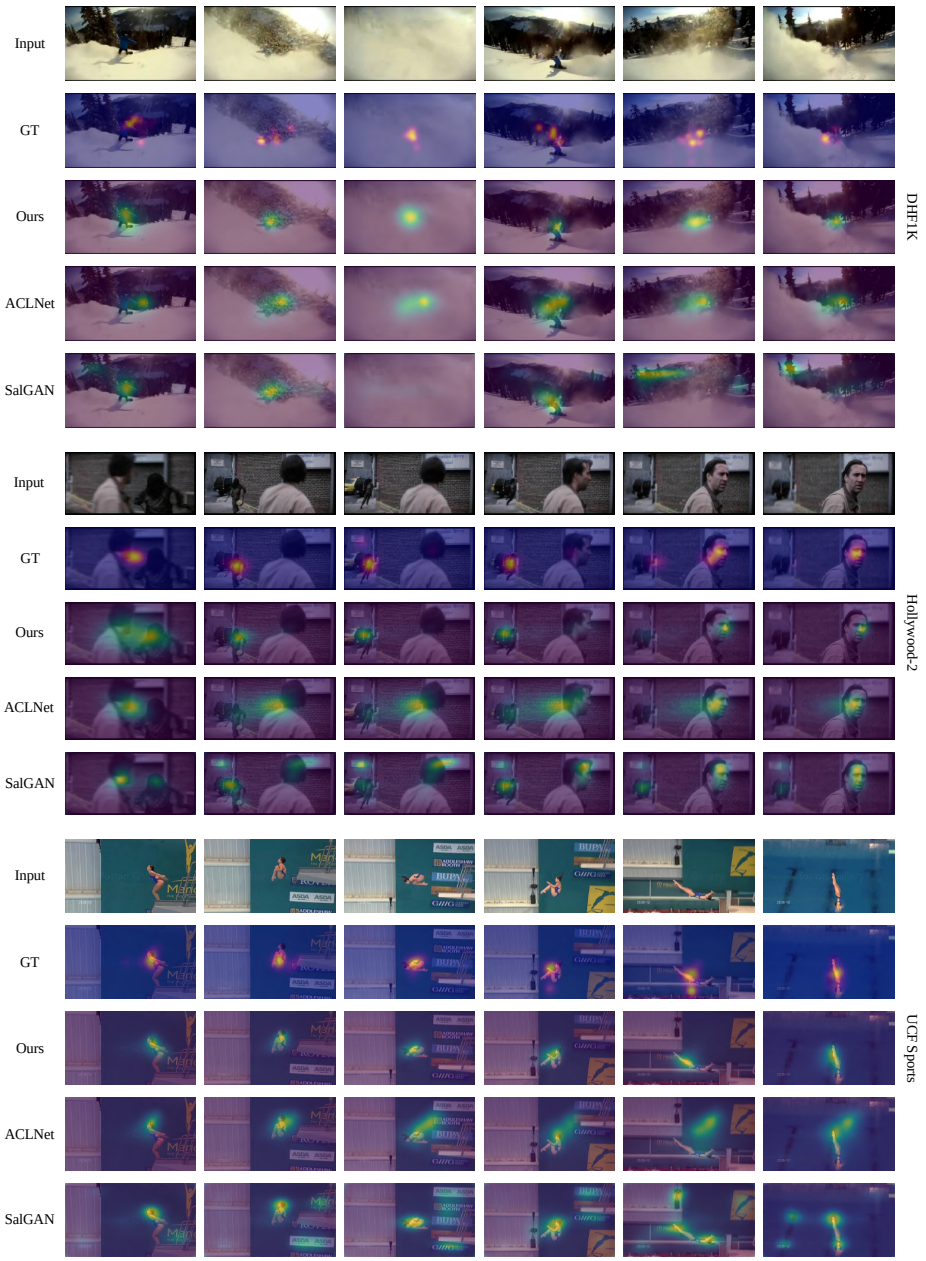

Fig. 1. Additional qualitative video saliency prediction results. Predictions of the proposed UNISAL model are compared to those of ACLNet [6] and SalGAN [4].

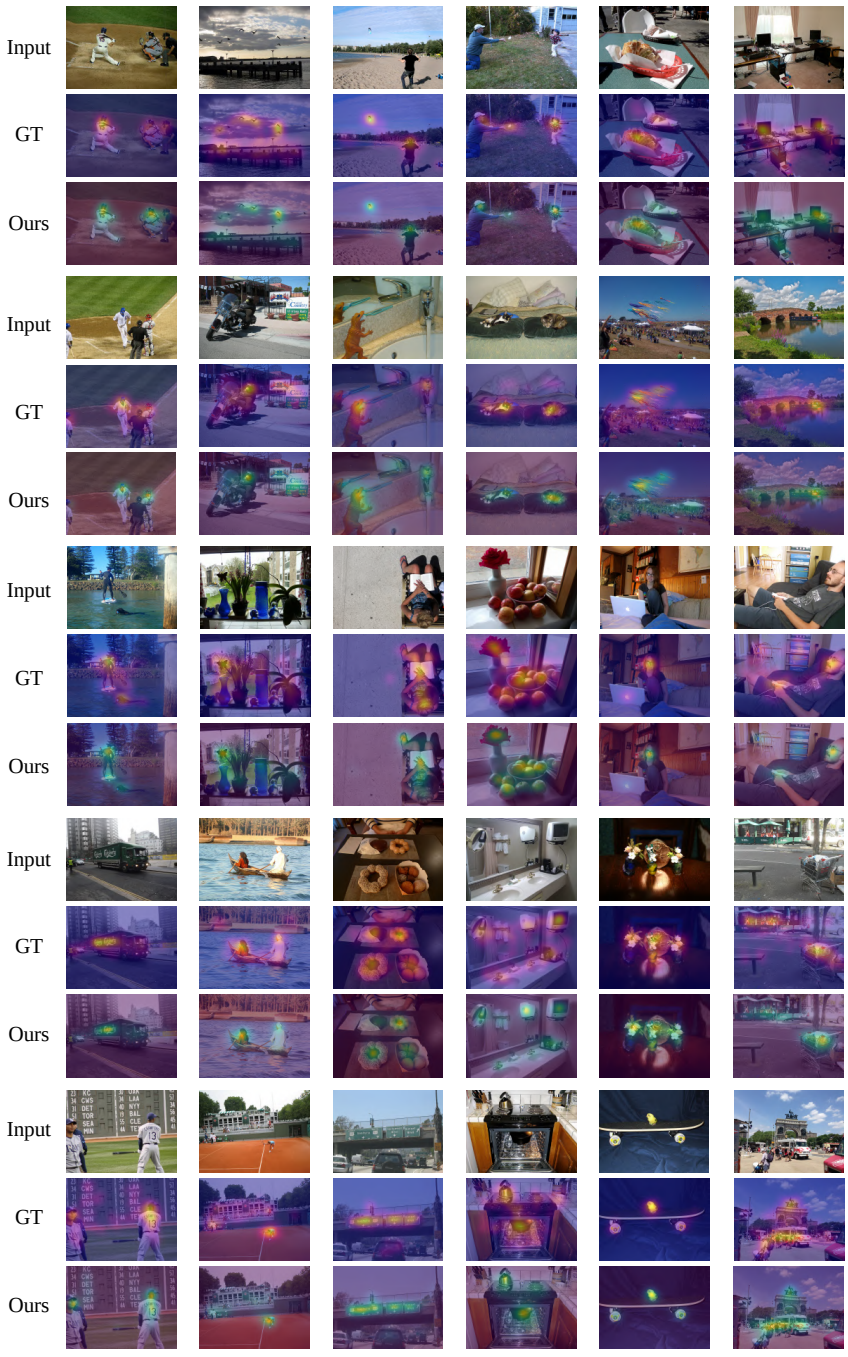

Fig. 2. Additional qualitative image saliency prediction results of the proposed UNISAL model for the SALICON dataset.

Fig. 3. Additional qualitative image saliency prediction results of the proposed UNISAL model for the MIT1003 dataset.

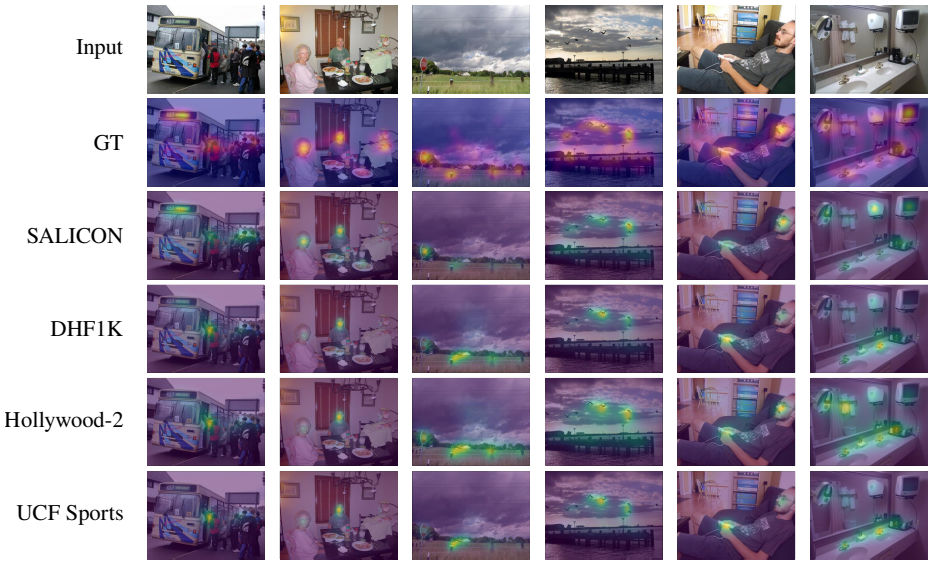

Fig. 4. Cross-domain predictions for SALICON. The images shown are drawn from the SALICON validation set. The predictions are generated with the same trained UNISAL model, but different domain-adaptive settings. The leftmost column shows the dataset whose modules were selected for the corresponding row.

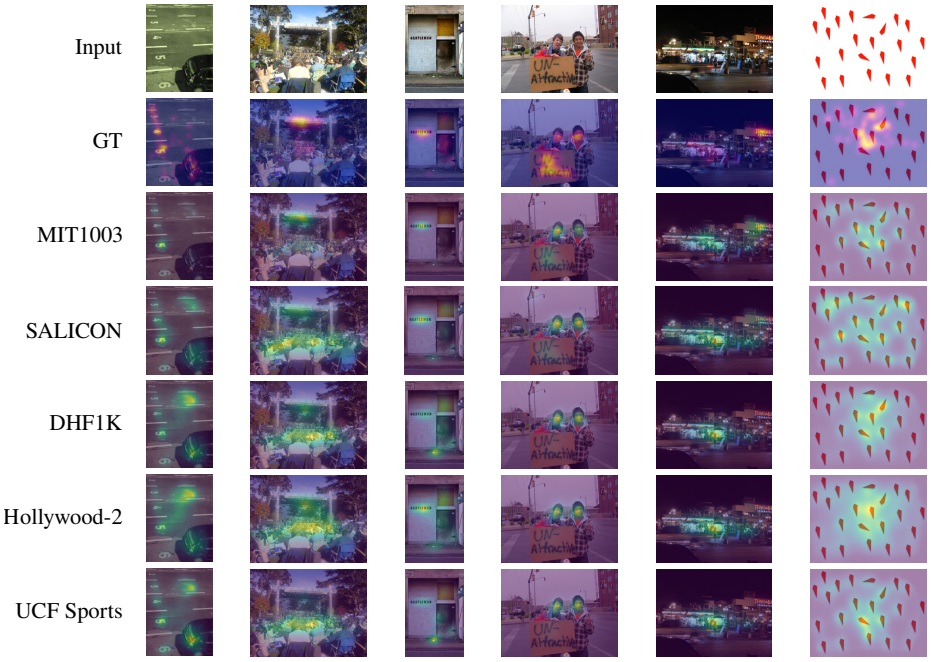

Fig. 5. Cross-domain predictions for MIT1003. The images shown are drawn from the MIT1003 dataset. The predictions are generated with the same trained UNISAL model, but different domain-adaptive settings. The leftmost column shows the dataset whose modules were selected for the corresponding row. MIT1003 denotes the SALICON-specific setting which was fine-tuned on MIT1003 samples.

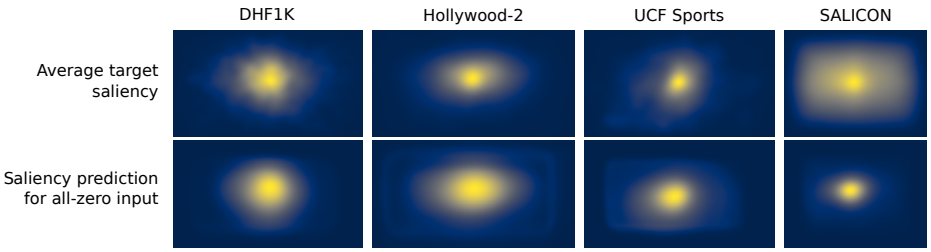

Fig. 6. Saliency targets center biases vs. learned biases. The upper row shows the average across all target training saliency maps for each dataset. The lower row shows the prediction of the model for an all-zero input, for different domain-adaptive settings.