Learning from Unique Perspectives: User-aware Saliency Modeling (Supplementary Materials)

Shi Chen*
University of Minnesota
chen4595@umn.edu

Nachiappan Valliappan
Google Research
nac@google.com

Shaolei Shen
Google
shaoleis@google.com

Xinyu Ye
Google
yexinyu@google.com

Kai Kohlhoff
Google Research
kohlhoff@google.com

Junfeng He †
Google Research
junfenghe@google.com

These supplementary materials consist of additional experimental results and analyses that demonstrate the effectiveness of our proposed method:

1. We evaluate our method on a recently introduced Webpage Saliency dataset [1] (Section 1).
2. We demonstrate the generalizability of our method by experimenting with an additional model, *i.e.,* SimpleNet [9] (Section 2).
3. We provide a comprehensive analysis on the effects of the inter-user agreement on capturing variability of attention (Section 3).
4. We provide supplementary results on leveraging single-user ensemble for general saliency prediction (Section 4).
5. We perform an ablation study to investigate the effectiveness of different components within our method (Section 5).
6. We present results with additional evaluation metrics to complement experiments in the main paper (Section 6).
7. We provide the detailed definition of training objectives used by our method (Section 7).

1. Results on Webpage Saliency Dataset

Comparative results in the main paper demonstrate the effectiveness of our method across diverse visual stimuli, including naturalistic images (OSIE [12]) and web pages (FiWI [11]). To further highlight its advantages, we carry out experiments on a recently introduced Webpage Saliency dataset [1] with eye-tracking data collected from 41 users on 450 web pages. According to the results in Table 1 and Table 2, our method leads to improvements consistent with those reported in the main paper, and is advantageous in both user-aware and general saliency prediction.

Table 1. User-aware saliency results on Webpage Saliency [1].

	K=1		K=3		K=5	
	NSS	CC	NSS	CC	NSS	CC
EML-Net	1.550	0.430	1.542	0.604	1.540	0.678
Ours*	1.622	0.438	1.592	0.618	1.590	0.698
Ours	1.669	0.446	1.607	0.626	1.602	0.703

Table 2. General saliency results on Webpage Saliency [1].

	NSS	KLD	CC	AUC-Judd	sAUC
EML-Net	1.538	0.284	0.800	0.838	0.723
Ours	1.590	0.251	0.827	0.842	0.749

2. Experimenting with SimpleNet

In addition to showing the generalizability of our method across different visual stimuli, we are also interested in validating its robustness to different architectural designs. To complement our results in the main paper, which are obtained by incorporating our method with EML-Net [5], we further experiment with a different model. We choose SimpleNet [9], which is a computationally efficient yet highly accurate model. As shown in Table 3 and Table 4, our method is able to outperform its counterparts without considering visual preferences (*i.e.,* SimpleNet [9]) or the composition of attention (*i.e.,* Ours*). The results agree with

*Work done during an internship at Google Research.
†Corresponding author

Table 3. User-aware saliency prediction results on FiWI dataset with SimpleNet.

	K=1		K=3		K=5	
	NSS	CC	NSS	CC	NSS	CC
SimpleNet	1.515	0.312	1.499	0.454	1.518	0.519
Ours*	1.603	0.314	1.352	0.404	1.507	0.516
Ours	1.738	0.341	1.634	0.489	1.634	0.558

Table 4. General saliency prediction results on FiWI dataset with SimpleNet.

	NSS	KLD	SIM	CC	AUC
SimpleNet	1.501	1.317	0.556	0.598	0.821
Ours	1.597	0.682	0.577	0.639	0.838

those in the main paper, and serve as supporting evidence for validating the generalizability of our method.

3. How Does Inter-user Agreement Affect User-aware Saliency Modeling?

Visual attention is driven by both users' visual preferences and the properties of visual stimuli. As a result, despite the variability in users' attention deployments, there exists a considerable agreement between the regions of interest by most users. As shown in Table 5, attention maps aggregated from K randomly selected users tend to have a relatively high correlation with those for all users. The inter-user agreement creates difficulties for models to understand the diversity of visual behaviors, and results in a shortcut where predicting an attention map averaged across all users is able to provide reasonable performance on user-aware saliency prediction. Our results in Table 6 show that, for our model without the proposed learning method (Ours*), leveraging an all-one user mask as input (+All-one) outperforms its counterpart with the correct presence of users (*i.e.,* only considering attention of the selected users) on CC scores. The observation implies that the model lacks the capability to encode the users' visual preferences with their corresponding personalized filters. We overcome the issue with the progressive learning method that enables the model to bridge users' preferences with saliency driven by visual stimuli (Ours). Comparative results show that it plays an important role in addressing the shortcut (using the correct instead of all-one user mask has considerably better results), which also leads to enhanced performance on user-aware saliency modeling (see our main paper).

Table 5. Correlation between attention of K and all users on FiWI dataset

Number of users	NSS	KLD	SIM	CC	AUC
K=1	1.548	8.716	0.368	0.522	0.732
K=3	2.168	3.547	0.578	0.731	0.830
K=5	2.471	1.868	0.683	0.834	0.873

Table 6. Comparative results between models using adaptive and all-one user masks on FiWI dataset. Note that the experiments are run with a different (random) selection of users from those in the main paper, which causes certain inconsistency.

	K=1		K=3		K=5	
	NSS	CC	NSS	CC	NSS	CC
Ours*+All-one	1.812	0.372	1.757	0.531	1.773	0.610
Ours*	1.860	0.362	1.783	0.530	1.784	0.609
Ours+All-one	1.823	0.374	1.776	0.537	1.790	0.616
Ours	1.987	0.388	1.821	0.541	1.815	0.620

Table 7. Comparative results on general saliency prediction for FiWI and OSIE datasets. For FiWI, we report the results for 5-fold cross-validation.

		NSS	KLD	SIM	CC	AUC
FiWI	EML-Net [5]	1.686	0.589	0.591	0.674	0.848
	Single-user	1.719	0.578	0.595	0.687	0.850
	Ours	**1.777**	**0.547**	**0.616**	**0.709**	**0.854**
OSIE	EML-Net [5]	1.737	0.537	0.619	0.717	0.854
	Single-user	1.629	0.606	0.588	0.687	0.842
	Ours	**1.840**	**0.506**	**0.652**	**0.761**	**0.860**

4. Does Single-user Ensemble Help General Saliency Prediction?

In the main paper, we show that our method, with a novel model architecture and a principled learning paradigm, is able to outperform an ensemble of single-user models (*i.e.,* Single-user, training one model for each individual user) in capturing the variability of users' attention. In this section, we further demonstrate its effectiveness by comparing them on the general saliency prediction tasks. As reported in Table 7, the additive ensemble either brings negligible improvements (FiWI with complete users), or leads to a visible drop of accuracy (OSIE with sparse user annotations). The results highlights the advantages of our approach in enhancing the predictive power on both user-aware and general saliency prediction tasks, and overcoming the issues of incomplete user annotations.

Table 8. Ablation results for user-aware saliency on FiWI dataset.

	K=1		K=3		K=5	
	NSS	CC	NSS	CC	NSS	CC
EML-Net [5]	1.481	0.307	1.506	0.454	1.498	0.519
Ours* w/o agg	1.912	0.374	1.816	0.539	1.785	0.613
Ours w/o agg	2.056	**0.399**	1.816	0.539	1.763	0.606
Ours w/o per-user	1.777	0.344	1.787	0.529	1.767	0.604
Ours	**2.059**	0.392	**1.829**	**0.540**	**1.815**	**0.620**

5. Ablation Study on Method Design

This section provides an ablation study on the contributions of different components in the proposed method. Specifically, we carry out experiments on three variants of our method, including (1) Ours* w/o agg, which resembles a multi-task learning model that shares the same architecture as our model but does not use progressive learning and is only optimized on per-user supervision (*i.e.,* the 2^{nd} term in equation (5) of the main paper), (2) Ours w/o agg, which is our model with progressive learning but only per-user optimization, and (3) Ours w/o per-user, which is our model with progressive learning but not per-user optimization.

We draw two major observations from the comparative results reported in Table 8, the experimental settings are the same as those discussed in the main paper:

- **Personalized filters play a key role in capturing visual preferences.** Compared to the user-agnostic EML-Net [5], Ours* w/o agg shows significant increases in all evaluation settings. The results verify the importance of leveraging our proposed personalized filters to encode discriminative preferences of different users.

- **Bridging users' preferences with the overall attention is important for user-aware saliency modeling.** Experimental results indicate that dropping supervision on either intermediate per-user predictions or the final attention output leads to a visible drop of performance. They demonstrate the integral design of our method, and more importantly, highlight the need to jointly consider the visual preferences of individual users and the composition of their attention patterns. With the consideration of both factors, our full method achieves overall the best results.

6. Results with Additional Metrics

To complement our results reported in the main paper (*e.g.,* user-aware saliency evaluated with Normalized Scanpath Saliency (NSS) [8] and Correlation Coefficient (CC) [7]), in this section, we present results with two additional metrics (*i.e.,* Similarity (SIM) [10] and AUC [3]). Comparative results reported in Table 9 and Table 10 are consistent with the observations discussed in the main paper, showing the advantages of our method under various evaluation settings.

Table 9. User-aware saliency results on FiWI dataset with additional metrics.

	K=1		K=3		K=5	
	SIM	AUC	SIM	AUC	SIM	AUC
EML-Net	0.261	0.820	0.413	0.821	0.479	0.821
Single-user	0.245	0.819	0.401	0.834	0.469	0.836
Ours*	0.299	0.863	0.452	0.856	0.521	0.856
Ours	**0.306**	**0.867**	**0.458**	**0.856**	**0.528**	**0.857**

Table 10. General saliency results on FiWI dataset with additional metrics.

	SIM	AUC
EML-Net	0.591	0.845
EML-Net+SALICON	0.602	0.846
Ours	**0.608**	**0.849**

7. Definition of Objective Function

Following [2, 4], we leverage a linear combination of saliency evaluation metrics as our loss function (*i.e.,* L_{sal} in the main paper). The function takes into account three popular metrics, including Normalized Scanpath Saliency (NSS) [8], Correlation Coefficient (CC) [7], and KL-Divergence (KLD) [6]:

$$L_{sal} = \alpha \cdot NSS(S, Fix) + \beta \cdot CC(S, Sal) + \gamma \cdot KLD(S, Sal) \quad (1)$$

where S is the predicted attention map, Fix and Sal are the ground truth fixation and saliency map, respectively. $\alpha = -1$, $\beta = -2$, and $\gamma = 10$ are balancing factors defined according to [2].

References

[1] Souradeep Chakraborty, Zijun Wei, Conor Kelton, Seoyoung Ahn, Aruna Balasubramanian, Gregory J. Zelinsky, and Dimitris Samaras. Predicting visual attention in graphic design documents. *IEEE Transactions on Multimedia*, pages 1–1, 2022. 1

[2] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018. 3

[3] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. 3

[4] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV*, pages 262–270, December 2015. 3

[5] Sen Jia and Neil D.B. Bruce. Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing*, 95:103887, 2020. 1, 2, 3

[6] Solomon Kullback. *Information Theory and Statistics*. Wiley, 1959. 3

[7] Olivier Le Meur, Patrick Le Callet, and Dominique Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47(19):2483 – 2498, 2007. 3

[8] Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397 – 2416, 2005. 3

[9] Navyasri Reddy, Samyak Jain, Pradeep Yarlagadda, and Vineet Gandhi. Tidying deep saliency prediction architectures. In *IROS*, 2020. 1

[10] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121, 2000. 3

[11] Chengyao Shen and Qi Zhao. Webpage saliency. In *ECCV*, pages 33–46, 2014. 1

[12] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of Vision*, 14(1):1–20, 2014. 1