SUM: Saliency Unification through Mamba for Visual Attention Modeling

Supplementary Material

Alireza Hosseini[*,1] Amirhossein Kazerouni[*,2,3,4] Saeed Akhavan [1]
Michael Brudno [2,3,4] Babak Taati [2,3,4]
[1] University of Tehran [2] University of Toronto [3] Vector Institute
[4] University Health Network
{arhosseini77, s.akhavan}@ut.ac.ir, {amirhossein, brudno}@cs.toronto.edu
babak.taati@uhn.ca

A. Experimental Results

A.1. Impact of different loss combinations

We examined how various combinations of loss metrics affected the validation performance of the model in Table 1. In addition, to provide additional details about the coefficients used for each loss combination, we conducted several experiments to determine the optimal coefficients for each combination. The best coefficients for each combination are depicted in Table 2.

A.2. More visualization results

We have included an additional visualization of SUM's predictions in Figure 1. Compared to ground truths, SUM consistently delivers accurate predictions across various image types and datasets, underscoring its robustness and versatility in visual saliency modeling. Moreover, to further validate the robustness of our proposed method, we conducted comparative analyses using publicly available datasets that had not been previously seen, as detailed in Table Table 3. The performance, as depicted in Figure 2, notably remains consistent when applied to new and previously unseen datasets. This suggests that SUM adeptly identifies and highlights the salient features in images, maintaining close alignment with the ground truth data. Therefore, SUM can be reliably utilized in diverse real-world applications where accuracy in visual recognition is critical.

References

[1] Hani Alers, Hantao Liu, Judith Redi, and Ingrid Heynderickx. Studying the effect of optimizing the image quality in saliency regions at the expense of background content. In *Image Quality and System Performance VII*, volume 7529, pages 59–67. SPIE, 2010. 2, 4

[2] Neil Bruce and John Tsotsos. Attention based on information maximization. *Journal of Vision*, 7(9):950–950, 2007. 2, 4

[3] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015. 2

[4] Hantao Liu and Ingrid Heynderickx. Studying the added value of visual attention in objective image quality metrics based on eye movement data. In *2009 16th IEEE international conference on image processing (ICIP)*, pages 3097–3100. IEEE, 2009. 2, 4

[5] Chengyao Shen and Qi Zhao. Webpage saliency. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 33–46. Springer, 2014. 2, 4

Table 1. Evaluation of different combinations of loss functions on model performance.

Loss Functions					Avg. Performance on Salicon [3]					Avg. Performance Across All Datasets				
KL	CC	SIM	NSS	MSE	CC ↑	KLD ↓	NSS ↑	SIM ↑	$\mathcal{F}_{\text{Score}}$ ↑	CC ↑	KLD ↓	NSS ↑	SIM ↑	$\mathcal{F}_{\text{Score}}$ ↑
✓	✗	✗	✗	✗	0.910	0.189	1.908	0.805	2.797	0.85	0.465	2.498	0.723	2.386
✗	✓	✗	✗	✗	0.907	0.732	1.926	0.787	1.634	0.851	1.08	2.532	0.7	1.218
✗	✗	✓	✗	✗	**0.911**	0.447	1.91	**0.807**	2.391	0.85	0.747	2.469	**0.728**	1.917
✗	✗	✗	✓	✗	0.834	0.765	**2.044**	0.721	0	0.804	1.072	2.614	0.658	-0.079
✗	✗	✗	✗	✓	0.909	0.234	1.919	0.803	2.696	0.846	0.525	2.479	0.719	2.089
✓	✗	✓	✗	✗	**0.911**	0.196	1.928	0.806	2.833	**0.852**	0.465	2.337	**0.728**	1.972
✓	✗	✗	✓	✗	0.892	0.199	2.029	0.792	2.537	0.841	0.467	2.594	0.712	2.353
✓	✓	✗	✗	✗	**0.911**	**0.185**	1.191	0.805	1.977	**0.852**	0.453	2.515	0.720	2.46
✓	✗	✗	✗	✓	0.909	0.192	1.917	0.802	2.755	0.851	0.456	2.504	0.723	2.441
✗	✓	✓	✗	✗	0.910	0.531	1.921	0.802	2.188	0.85	0.871	2.503	0.721	1.733
✓	✓	✓	✗	✗	0.909	0.198	1.920	0.803	2.759	**0.852**	0.464	2.527	0.726	2.568
✓	✗	✓	✗	✓	0.909	0.192	1.919	0.799	2.722	**0.852**	0.461	2.514	0.726	2.53
✓	✗	✗	✓	✓	0.887	0.208	2.038	0.788	2.421	0.830	0.472	**2.642**	0.711	2.259
✓	✓	✗	✗	✓	0.910	0.188	1.914	0.803	2.783	0.851	**0.447**	2.511	0.722	2.464
✓	✓	✓	✓	✗	0.907	0.198	1.989	0.803	2.815	0.850	0.466	2.614	0.725	2.794
✓	✓	✓	✗	✓	0.905	0.208	1.920	0.798	2.632	**0.852**	0.457	2.510	0.720	2.437
✓	✓	✓	✓	✓	0.909	0.192	1.981	0.804	**2.853**	**0.852**	0.450	2.602	0.726	**2.836**

Table 2. loss weighting coefficients λ_i ($i = 1, \ldots, 5$) as used in Table 1.

KL	CC	SIM	NSS	MSE
1	0	0	0	0
0	-1	0	0	0
0	0	-1	0	0
0	0	0	-1	0
0	0	0	0	1
10	0	-3	0	0
10	0	0	-3	0
10	-3	0	0	0
10	0	0	0	5
0	-2	0	-1	0
10	-2	-1	0	0
10	0	-3	0	5
10	0	0	-3	5
10	-3	0	0	5
10	-2	-1	-1	0
10	-2	-1	0	5
10	-2	-1	-1	5

Table 3. Details of unseen datasets used for quantitative analysis of SUM in Figure 2.

Dataset	Image domain	# Image	Image Resolution
Toronto [2]	Natural scene	120	681 × 511
TUD Image Quality Database 1 [4]	Natural scene	29	768 × 512
TUD Image Quality Database 2 [1]	Natural scene	160	600 × 600
FIWI [5]	Web page	149	1360 × 768

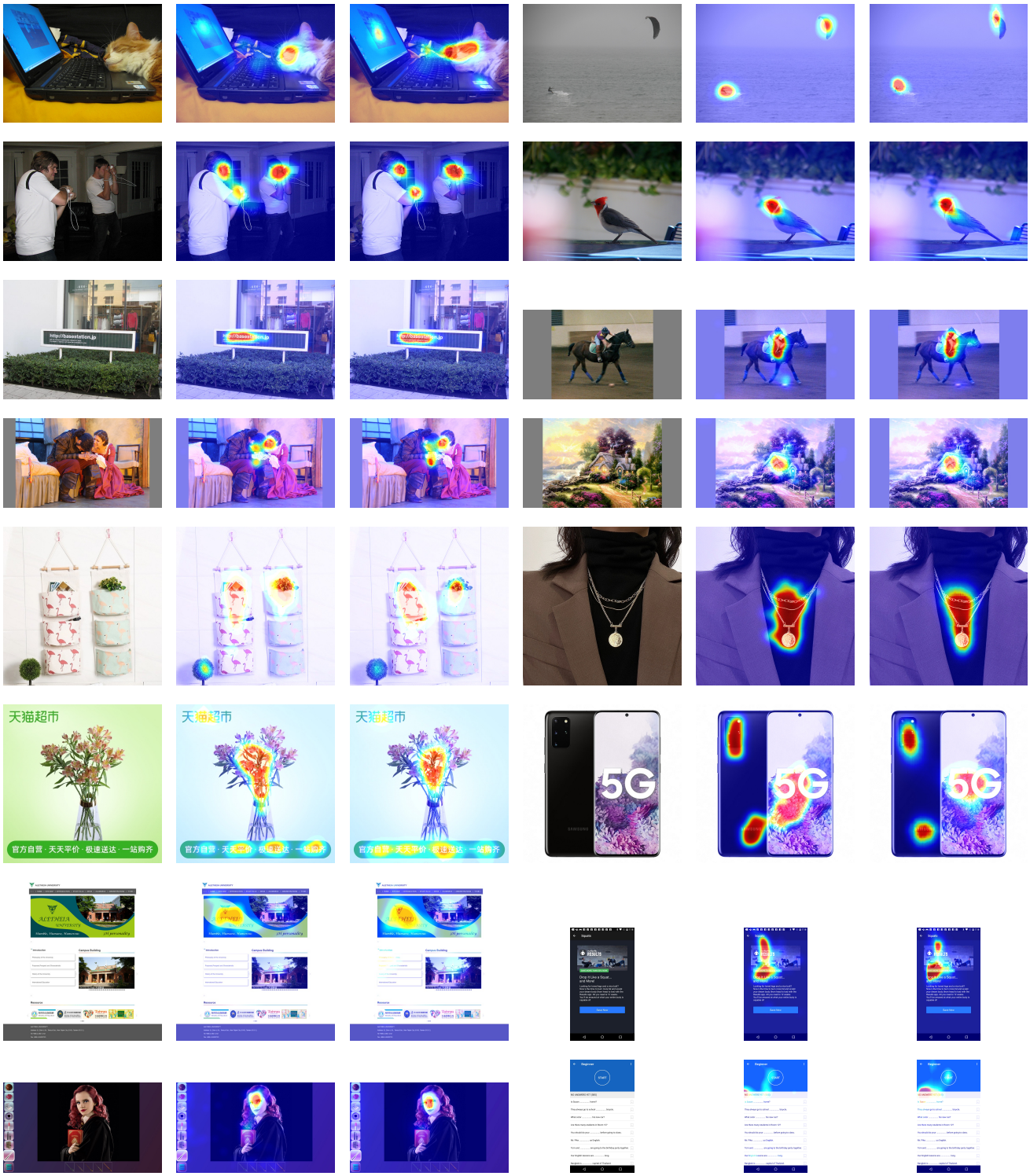
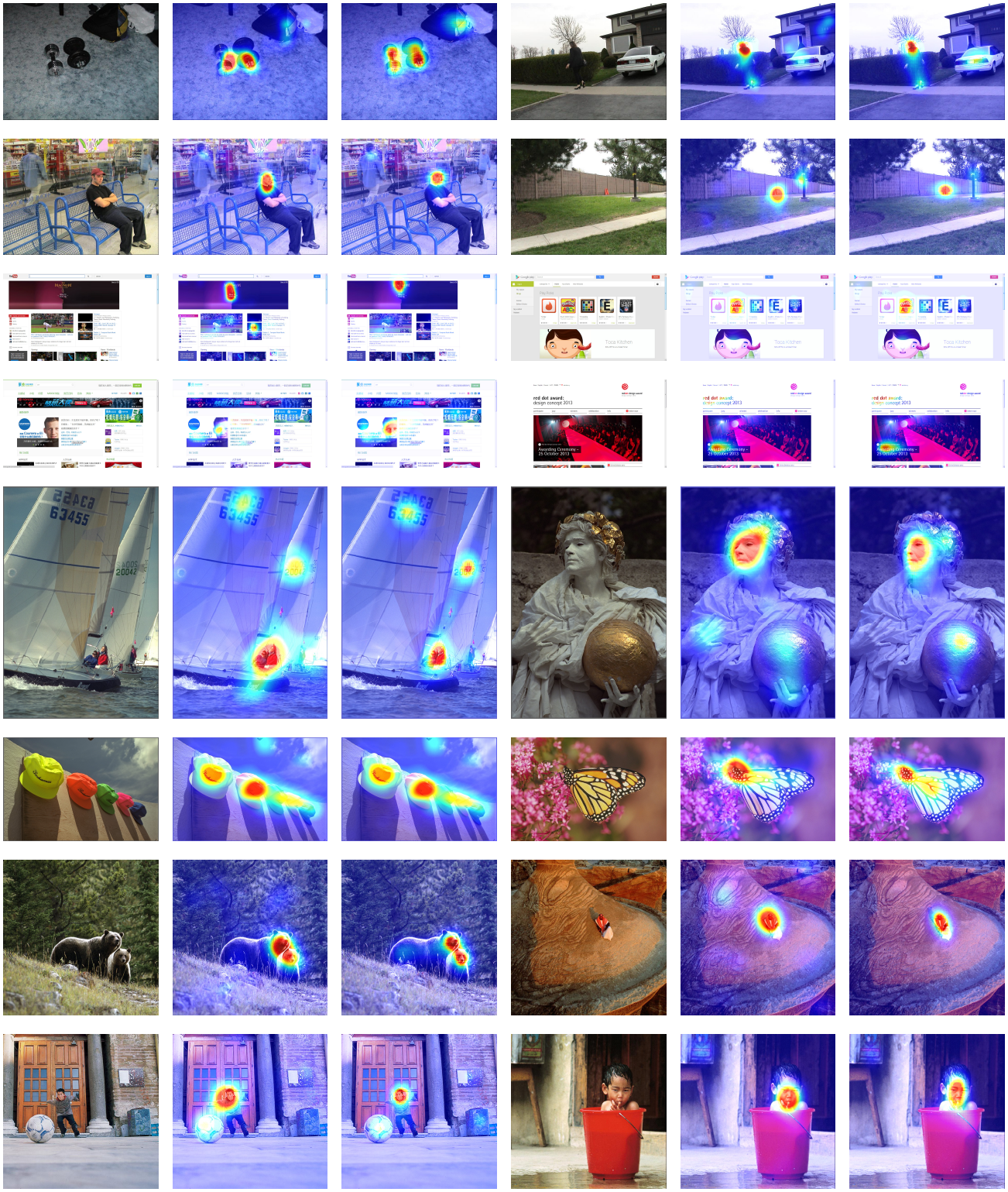


Figure 1. Visualizations of SUM's predictions across different datasets. The first and second rows depict Natural Scene-Mouse data, while the third and fourth rows showcase Natural Scene-Eye data. The fifth and sixth rows present E-commerce data, and the seventh and eighth rows display UI data.



Input Image | Ground Truth | **SUM** | Input Image | Ground Truth | **SUM**

Figure 2. Visualizations of SUM's predictions across different datasets. The first and second rows showcase the Toronto dataset [2], while the third and fourth rows present the FIWI dataset [5]. The fifth and sixth rows display data from the TUD Image Quality Database 1 [4], and the seventh and eighth rows exhibit data from the TUD Image Quality Database 2 [1].