Three-Stage Bidirectional Interaction Network for Efficient RGB-D Salient Object Detection (Supplementary Material)

Yang Wang and Yanqing Zhang✉

South China University of Technology, Guangzhou, China
202021045142@mail.scut.edu.cn
zyqcs@scut.edu.cn

Abstract. In this supplementary material, we explain more about the design of the three-stage bidirectional interaction (TBI) strategy and additionally analyze the effectiveness of the TBI strategy.

1 The Design of TBI

We employ bidirectional attention guidance (BAG) modules and bidirectional feature supplement (BFS) modules in the first two stages, respectively. Both the BAG module and the BFS module adopt a simple design. The sum of the parameters of all BAG modules and BFS modules in TBINet does not exceed 0.04M (the entire model parameter is 3.7M). The BAG module and the BFS module help the model achieve better saliency detection performance while slightly changing the model complexity. In addition, the adoption of shared network in the third stage reduces model parameters by more than 2.6M. The TBI strategy conducts appropriate interaction operations at specific stages, and it is a simple and effective cross-modality interaction strategy for RGB-D SOD.

Table 1. The additional effectiveness analyses of TBI strategy

Strategy		STERE		NJU2K		NLPR		SIP	
		$F_\beta^{\max}\uparrow$	MAE $\downarrow$	$F_\beta^{\max}\uparrow$	MAE $\downarrow$	$F_\beta^{\max}\uparrow$	MAE $\downarrow$	$F_\beta^{\max}\uparrow$	MAE $\downarrow$
Ours	S_F	.910	.034	.928	.029	.932	.018	.905	.041
	S_R	.906	.036	.922	.032	.925	.021	.896	.043
	S_D	.881	.047	.914	.035	.909	.025	.894	.047
w/o TBI	S_F	.906	.038	.919	.033	.929	.020	.891	.048
	S_R	.903	.038	.906	.039	.916	.022	.869	.056
	S_D	.711	.108	.859	.058	.856	.039	.865	.058

2 Additional Effectiveness Analysis of TBI

We completely remove the TBI strategy, denoted as 'w/o TBI'. We obtain the saliency maps S_F, S_R, and S_D generated by the three branches of the decoder

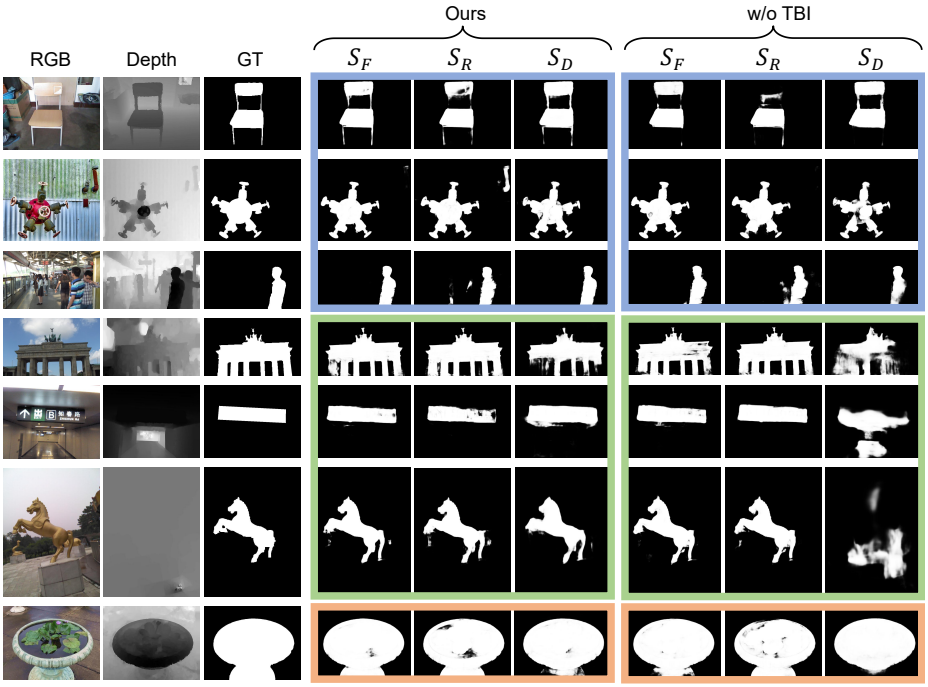

Fig. 1. Visual comparison of our model with the model without TBI. S_F, S_R and S_D denote the saliency maps output by the fusion branch, the RGB branch and the depth branch, respectively. 'w/o TBI' means without TBI strategy

to analyze the effect of the TBI strategy. The evaluation results on STERE [1], NJU2K [2], NLPR [3] and SIP [4] datasets are shown in Table 1. For S_F, S_R, and S_D, our method outperforms 'w/o TBI' by ($0.32\% \sim 1.50\%$, $0.002 \sim 0.007$), ($0.36\% \sim 3.18\%$, $0.001 \sim 0.013$), and ($3.44\% \sim 23.89\%$, $0.011 \sim 0.061$), respectively, for the metrics ($F_\beta^{\max}$ [5] and MAE [6]). This experiment shows the effectiveness of the TBI strategy. The model with TBI strategy performs better than the model without TBI strategy. It is worth noting that the performance gap between the two on S_D is huge. The reason is that depth maps are more likely to contain low-quality information than RGB images, and the TBI strategy can effectively supplement saliency-related information for the two modalities.

As shown in Fig. 1, we present some examples to demonstrate the role of the TBI strategy clearly. The 1^{st} to 3^{rd} rows show examples of RGB images with misleading information. In the 1^{st} row, there is a reflection phenomenon in the RGB image, which makes a part of the salient object indistinguishable from the background. The 2^{nd} and 3^{rd} rows show misleading objects in RGB images. Our method uses depth information to help locate the correct saliency regions. The 4^{th} to 6^{th} rows show examples of depth maps with misleading information. The S_D of 'w/o TBI' is entirely wrong, while the S_D of our method still locates

salient objects. Our method achieves promising results in suppressing low-quality depth information. The last row shows an example of cross-modality information complementarity. Part of the texture of the salient object in the RGB image is confused with the background, and the depth map cannot express the part of the salient object that is connected to the background. The 'w/o TBI' generates a flawed result. Our method combines the advantages of two modalities to obtain an accurate saliency map.

References

1. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 454–461. IEEE (2012)
2. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: 2014 IEEE international conference on image processing (ICIP). pp. 1115–1119. IEEE (2014)
3. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: Rgbd salient object detection: A benchmark and algorithms. In: European conference on computer vision. pp. 92–109. Springer (2014)
4. Fan, D.P., Lin, Z., Zhang, Z., Zhu, M., Cheng, M.M.: Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. IEEE Transactions on neural networks and learning systems **32**(5), 2075–2089 (2020)
5. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 1597–1604. IEEE (2009)
6. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 733–740. IEEE (2012)