RGB-D Salient Object Detection with Cross-Modality Modulation and Selection

Chongyi Li*[1], Runmin Cong*✉[2], Yongri Piao[3], Qianqian Xu[4], and Chen Change Loy[1]

[1] Nanyang Technological University, Singapore
[2] Beijing Jiaotong University, China
[3] Dalian University of Technology, China
[4] Institute of Computing Technology, Chinese Academy of Sciences, China
lichongyi25@gmail.com, rmcong@bjtu.edu.cn
yrpiao@dlut.edu.cn, xuqianqian@ict.ac.cn, ccloy@ntu.edu.sg
https://li-chongyi.github.io/Proj_ECCV20

Abstract. In this supplementary material, we first provide more visual results in Sec. 1, then analyze the side outputs of our network in Sec. 2, and finally compare the model sizes of different SOD methods in Sec. 3.

1 Visual Results

In this section, we provide more visual results of all the compared methods on the testing datasets in Fig. 1. In comparison, our method yields more complete, sharp, and edge-preserving saliency detection results, and effectively suppresses the cluttered backgrounds.

2 Side Outputs

In this section, we analyze the side outputs of our network. Since our network produces five saliency maps with a resolution ranging from 14×14 to 224×224 with a scale of 2, it can provide diverse choices based on salient object detection (SOD) accuracy and inference speed.

In some cases that require faster inference speed, we can perform early stopping on the inference and directly up-sample (such as by linear interpolation) the side output in the higher level to the same size of input RGB image as the final result. In this way, the inference time can be reduced when the SOD performance decreases accordingly as shown in Table 1. Visual examples of our side outputs in different levels are shown in Fig. 2.

* Equal contribution
✉ Corresponding author: *Runmin Cong* (rmcong@bjtu.edu.cn).

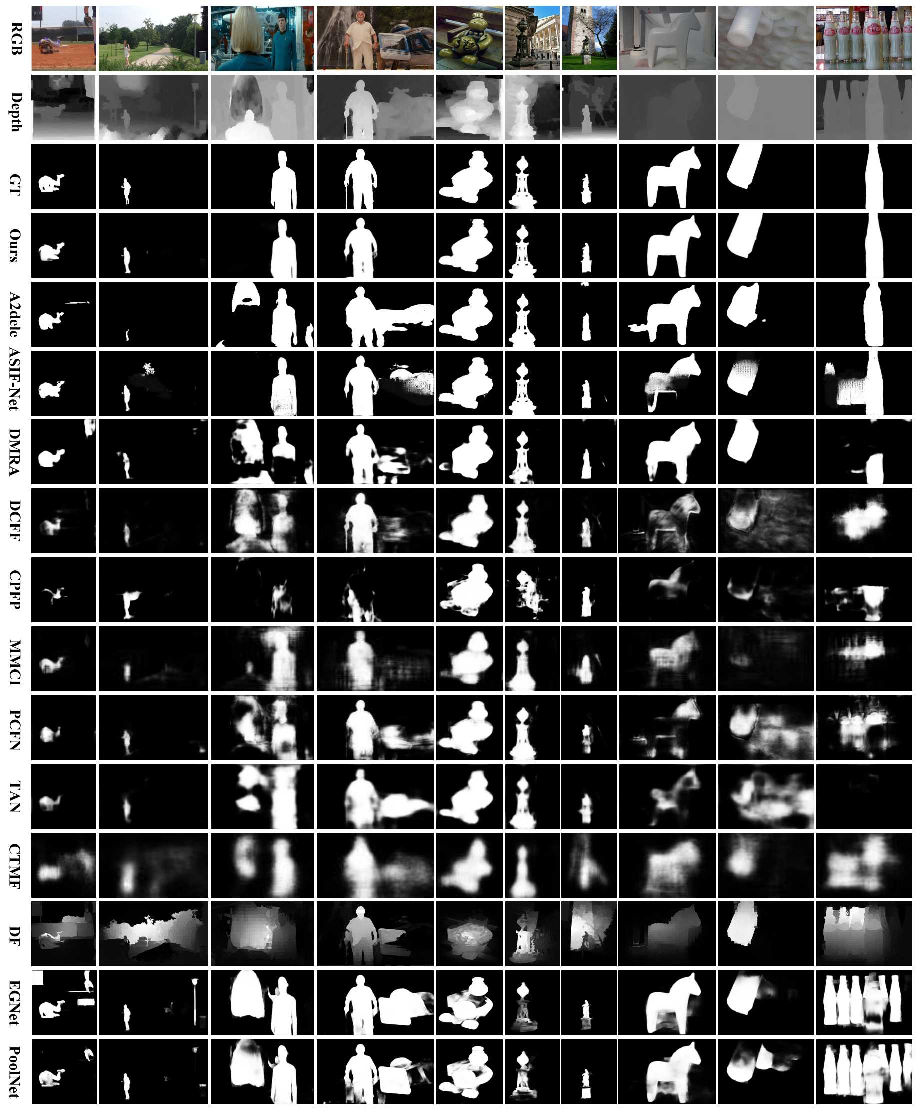

Fig. 1: Visual examples of different methods. From top to bottom are RGB images, the corresponding depth images, ground truth images, our results, the results of A2dele [13], ASIF-Net [9], DMRA [12], DCFF [4], CPFP [15], MMCI [5], PCFN [2], TAN [3], CTMF [8], DF [14], EGNet [16]), and PoolNet [10].

Table 1: Quantitative comparisons of side outputs in different levels of our network. "best competitor" represents the second best score under each metric in the main manuscript

Levels	STEREO Dataset [11]			DUT-Test Dataset [12]		
	F_β [1] ↑	MAE [6] ↓	S_m [7] ↑	F_β [1] ↑	MAE [6] ↓	S_m [7] ↑
level 1	**0.9084**	**0.0422**	**0.8895**	**0.9328**	**0.0366**	**0.8853**
level 2	0.9076	0.0442	0.8913	0.9319	0.0388	0.8866
level 3	0.9058	0.0504	0.8924	0.9296	0.0450	0.8894
level 4	0.8984	0.0642	0.8862	0.9212	0.0586	0.8843
level 5	0.8839	0.0909	0.8659	0.9057	0.0833	0.8667
best competitor	0.8997	0.0431	0.8778	0.9145	0.0426	0.8637

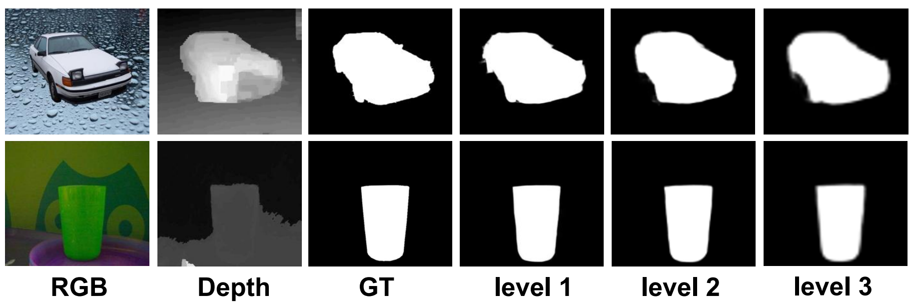

Fig. 2: Visual examples of the side outputs of our network. The sizes of outputs in level 2 and level 3 are up-sampled to the same size as the input RGB image by using linear interpolation.

As shown in Fig. 2 and Table 1, the outputs in level 2 and level 3 also achieve competitive SOD performance when they are compared with the output in level 1, but have faster inference speed. The inference speed of our network in level 1, level 2, and level 3 is 27 FPS, 31 FPS, and 38 FPS, respectively, for a pair of input RGB-D images with a size of 224×224. As presented in Table 1, compared with the "best competitor" among all the comparisons in the main manuscript, the scores of F-measure and S-measure of the output in level 3 are still higher on the STEREO dataset. Moreover, the scores of F-measure and S-measure of the output in level 4 are still superior on the DUT-Test dataset. In this paper, we treat the output in level 1 as the final result based on its more accurate and robust SOD performance, but have diverse choices by considering the balance of accuracy and inference speed.

3 Model Sizes

In this section, we compare the model sizes of different methods in Table 2. In this comparison, we discard the method of DF [14] because this method contains non-deep learning-based algorithm.

Table 2: The comparisons of model sizes of different methods (in MB)

Method	Ours	PoolNet [10]	EGNet [16]	CTMF [8]	PCFN [2]	MMCI [5]
Model Size	270.4	210.0	432.4	825.8	533.6	929.7
Method	TAN [3]	CPFP [15]	DCFF [4]	DMRA [12]	ASIF-Net [9]	A2dele [13]
Model Size	951.9	278.4	941.5	238.8	323.9	60.1

As presented in Table 2, our method has comparable model size with the state-of-the-art methods such as PoolNet [10], CPFP [15], and DMRA [12], and is more efficient than most compared methods such as EGNet [16], CTMF [8], PCFN [2], MMCI [5], TAN [3], DCFF [4], and ASIF-Net [9].

References

1. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: A benchmark. IEEE Trans. Image Process. **24**(12), 5706–5722 (2015)
2. Chen, H., Li, Y.: Progressively complementarity-aware fusion network for RGB-D salient object detection. In: CVPR. pp. 3051–3060 (2018)
3. Chen, H., Li, Y.: Three-stream attention-aware network for RGB-D salient object detection. IEEE Trans. Image Process. **28**(6), 2825–2835 (2019)
4. Chen, H., Li, Y., Su, D.: Discriminative cross-modal transfer learning and densely cross-level feedback fusion for RGB-D salient object detection. IEEE Trans. Cybern. pp. 1–13 (2019)
5. Chen, H., Li, Y., Su, D.: Multi-modal fusion network with multiscale multi-path and cross-modal interactions for RGB-D salient object detection. Pattern Recognit. **86**, 376–385 (2019)
6. Cong, R., Lei, J., Fu, H., Cheng, M.M., Lin, W., Huang, Q.: Review of visual saliency detectioin with comprehensive information. IEEE Trans. Circuits Syst. Video Technol **29**(10), 2941–2959 (2019)
7. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: ICCV. pp. 4548–4557 (2017)
8. Han, J., Chen, H., Liu, N., Yan, C., Li, X.: CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. IEEE Trans. Cybern. **48**(11), 3171–3183 (2018)
9. Li, C., Cong, R., Kwong, S., Hou, J., Fu, H., Zhu, G., Zhang, D., Huang, Q.: ASIF-Net: Attention steered interweave fusion network for RGBD salient object detection. IEEE Trans. Cybern. pp. 1–13 (2020)
10. Liu, J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: CVPR. pp. 3917–3926 (2019)

11. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: CVPR. pp. 454–461 (2012)
12. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection. In: ICCV. pp. 7254–7263 (2019)
13. Piao, Y., Rong, Z., Zhang, M., Ren, W., Lu, H.: A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In: CVPR. pp. 9060–9069 (2020)
14. Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., Yang, Q.: RGBD salient object detection via deep fusion. IEEE Trans. Image Process. **26**(5), 2274–2285 (2017)
15. Zhao, J., Cao, Y., Fan, D.P., Cheng, M.M., Li, X.Y., Zhang, L.: Contrast prior and fluid pyramid integration for RGBD salient object detection. In: CVPR. pp. 3927–3936 (2019)
16. Zhao, J., Liu, J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: EGNet: Edge guidance network for salient object detection. In: ICCV. pp. 8779–8788 (2019)