Supplementary Material
Suppress and Balance: A Simple Gated Network for Salient Object Detection

Xiaoqi Zhao[1†], Youwei Pang[1†], Lihe Zhang[1⋆], Huchuan Lu[1,2], and Lei Zhang[3,4]

[1] Dalian University of Technology, China
[2] Peng Cheng Laboratory
[3] Dept. of Computing, The Hong Kong Polytechnic University, China
[4] DAMO Academy, Alibaba Group
{zxq,lartpang}@mail.dlut.edu.cn, {zhanglihe,lhchuan}@dlut.edu.cn,
cslzhang@comppolyu.edu.hk

In this supplementary material, we expand our GateNet to other tasks including RGB-D Salient Object Detection (SOD) and Video Object Segmentation (VOS) to further demonstrate its effectiveness.

1 Network Architecture

Fig. 1 shows our proposed dual-branch gated FPN network for RGB-D SOD and VOS. Compared with the RGB SOD network, we only add an extra encoder to extract features of other modals such as depth or optical flow. This dual-branch GateNet is easy to follow and can be used as a new baseline.

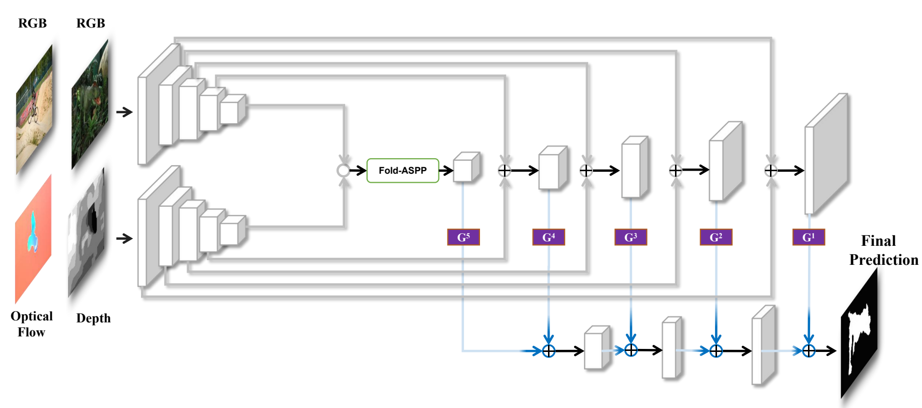

Fig. 1. Network pipeline.

⋆ Corresponding author.
† These authors contributed equally to this work.

2 RGB-D Salient object detection

2.1 Dataset

There are five main RGB-D SOD datasets which are **NJUD** [11], **RGBD135** [4] **NLPR** [14], **SSD** [28] and **SIP** [9]. We adopt the same splitting way as [1,3,10,26,16] to guarantee a fair comparison. We split 1,485 samples from NJUD and 700 samples from NLPR for traing a new model. The remaining images in these two datasets and other three datasets are all for testing to verify the generalization ability of saliency models.

2.2 Evaluation Metrics

We adopt several metrics widely used in RGB-D SOD for quantitative evaluation: F-measure score, mean absolute error (MAE, $\mathcal{M}$), the recently released S-measure (S_m) [7] and E-measure (E_m) [8] scores. The lower value is better for the MAE and higher is better for others.

2.3 Comparison with State-of-the-art Results

The performance of the proposed model is compared with ten state-of-the-art approaches on five benchmark datasets, including the DES [4], DCMC [5], CDCP [29], DF [17], CTMF [10], PCA [1], MMCI [3], TANet [2], CPFP [26] and DMRA [16]. For fair comparisons, all the saliency maps of these methods are directly provided by authors or computed by their released codes. And we take the VGG-16 as the backbone for each stream. Tab. 1 shows performance comparisons in terms of the maximum F-measure, mean F-measure, weighted F-measure, S-measure, E-measure and MAE scores. It can be seen that our GateNet is very competitive. We believe that future works based on GateNet can further improve performance and easily become the state-of-the-art RGB-D SOD model.

3 Video Object Segmentation

According to whether the mask of the first frame of the video is provided during the test, video object segmentation (vos) can be divided into zero-shot vos and one-shot vos. In this paper, we mainly use the dual-branch GateNet structure as shown in Fig. 1 for zero-shot vos.

3.1 Dataset and Metrics

DAVIS-16 [15] is one of the most popular benchmark datasets for video object segmentation tasks. It consists of 50 high-quality video sequences (30 for training and 20 for validation) in total. **Youtube-VOS** [24] is the latest large-scale dataset for the video object segmentation that consists of 4,453 videos annotated with multiple objects. We follow the training strategy as AGS [23] and MAT-Net [27] to obtain 13,438 training images in total. For quantitative evaluation, we adopt two metrics, namely region similarity $\mathcal{J}$ and boundary accuracy $\mathcal{F}$.

3.2 Comparison with State-of-the-art Results

The performance of the proposed model is compared with ten state-of-the-art approaches on the DAVIS-16 dataset, including the LVO [21], ARP [12], PDB [19], LSMO [22], MotAdapt [18], EPO [6], AGS [23], COSNet [13], AnDiff [25] and MATNet [27]. We follow most methods [27,25,13,22] to take the ResNet-101 as the backbone. Tab. 2 shows performance comparisons in terms of the $\mathcal{J}$ and $\mathcal{F}$. It should be noted that our method only performs feature extraction on the optical flow map generated by PWCNet [20] in order to supplement the motion information of the current frame. Without adding more cross-modal fusion techniques, or using other tracking or detection models, our GateNet can achieve competitive performance with most zero-shot vos methods.

Table 1. Quantitative comparison. ↑ and ↓ indicate that the larger and smaller scores are better, respectively. Among the CNN-based methods, the best results are shown in **red**. The subscript in each model name is the publication year.

Dataset	Metric	Traditional Methods			CNNs-Based Models							
		DES_{14} [4]	$DCMC_{16}$ [5]	$CDCP_{17}$ [29]	DF_{17} [17]	$CTMF_{18}$ [10]	$PCANet_{18}$ [1]	$MMCI_{19}$ [3]	$TANet_{19}$ [2]	$CPFP_{19}$ [26]	$DMRA_{19}$ [16]	GateNet Ours
SSD [28]	$F_\beta^{max}\uparrow$	0.260	0.750	0.576	0.763	0.755	0.844	0.823	0.835	0.801	0.858	**0.868**
	$F_\beta^{mean}\uparrow$	0.073	0.684	0.524	0.709	0.709	0.786	0.748	0.767	0.726	0.821	**0.822**
	$F_\beta^{w}\uparrow$	0.172	0.480	0.429	0.536	0.622	0.733	0.662	0.727	0.709	**0.787**	0.785
	$S_m\uparrow$	0.341	0.706	0.603	0.741	0.776	0.842	0.813	0.839	0.807	0.856	**0.870**
	$E_m\uparrow$	0.475	0.790	0.714	0.801	0.838	0.890	0.860	0.886	0.832	0.898	**0.901**
	$\mathcal{M}\downarrow$	0.500	0.168	0.219	0.151	0.100	0.063	0.082	0.063	0.082	0.059	**0.055**
NJUD [11]	$F_\beta^{max}\uparrow$	0.328	0.769	0.661	0.789	0.857	0.888	0.868	0.888	0.890	0.896	**0.914**
	$F_\beta^{mean}\uparrow$	0.165	0.715	0.618	0.744	0.788	0.844	0.813	0.844	0.837	0.871	**0.879**
	$F_\beta^{w}\uparrow$	0.234	0.497	0.510	0.545	0.720	0.803	0.739	0.805	0.828	0.847	**0.849**
	$S_m\uparrow$	0.413	0.703	0.672	0.735	0.849	0.877	0.859	0.878	0.878	0.885	**0.902**
	$E_m\uparrow$	0.491	0.796	0.751	0.818	0.866	0.909	0.882	0.909	0.900	0.920	**0.922**
	$\mathcal{M}\downarrow$	0.448	0.167	0.182	0.151	0.085	0.059	0.079	0.061	0.053	0.051	**0.047**
RGBD135 [4]	$F_\beta^{max}\uparrow$	0.800	0.311	0.651	0.625	0.865	0.842	0.839	0.853	0.882	0.906	**0.919**
	$F_\beta^{mean}\uparrow$	0.695	0.234	0.594	0.573	0.778	0.774	0.762	0.795	0.829	0.867	**0.891**
	$F_\beta^{w}\uparrow$	0.301	0.169	0.478	0.392	0.687	0.711	0.650	0.740	0.787	**0.843**	0.838
	$S_m\uparrow$	0.632	0.469	0.709	0.685	0.863	0.843	0.848	0.858	0.872	0.899	**0.905**
	$E_m\uparrow$	0.817	0.676	0.810	0.806	0.911	0.912	0.904	0.919	0.927	0.944	**0.966**
	$\mathcal{M}\downarrow$	0.289	0.196	0.120	0.131	0.055	0.050	0.065	0.046	0.038	**0.030**	**0.030**
NLPR [14]	$F_\beta^{max}\uparrow$	0.695	0.413	0.687	0.752	0.841	0.864	0.841	0.876	0.884	0.888	**0.904**
	$F_\beta^{mean}\uparrow$	0.583	0.328	0.592	0.683	0.724	0.795	0.730	0.796	0.818	**0.855**	0.854
	$F_\beta^{w}\uparrow$	0.254	0.259	0.501	0.516	0.679	0.762	0.676	0.780	0.807	**0.840**	0.838
	$S_m\uparrow$	0.582	0.550	0.724	0.769	0.860	0.874	0.856	0.886	0.884	0.898	**0.910**
	$E_m\uparrow$	0.760	0.685	0.786	0.840	0.869	0.916	0.872	0.916	0.920	**0.942**	**0.942**
	$\mathcal{M}\downarrow$	0.301	0.196	0.115	0.100	0.056	0.044	0.059	0.041	0.038	**0.031**	0.032
SIP [9]	$F_\beta^{max}\uparrow$	0.720	0.680	0.544	0.704	0.720	0.861	0.840	0.851	0.870	0.847	**0.894**
	$F_\beta^{mean}\uparrow$	0.644	0.645	0.495	0.673	0.684	0.825	0.795	0.809	0.819	0.815	**0.856**
	$F_\beta^{w}\uparrow$	0.342	0.414	0.397	0.406	0.535	0.768	0.712	0.748	0.788	0.734	**0.810**
	$S_m\uparrow$	0.616	0.683	0.595	0.653	0.716	0.842	0.833	0.835	0.850	0.800	**0.874**
	$E_m\uparrow$	0.751	0.787	0.722	0.794	0.824	0.900	0.886	0.894	0.899	0.858	**0.914**
	$\mathcal{M}\downarrow$	0.298	0.186	0.224	0.185	0.139	0.071	0.086	0.075	0.064	0.088	**0.057**

Table 2. Quantitative comparison of Zero-shot VOS methods on the DAVIS-16 validation set. ↑ and ↓ indicate that the larger and smaller scores are better, respectively. The best results are shown in **red**. The subscript in each model name is the publication year.

Metric		LVO_{17} [21]	ARP_{17} [12]	PDB_{18} [19]	$LSMO_{19}$ [22]	$MotAdapt_{19}$ [18]	EPO_{20} [6]	AGS_{19} [23]	$COSNet_{19}$ [13]	$AnDiff_{19}$ [25]	$MATNet_{20}$ [27]	GateNet Ours
$\mathcal{J}$	Mean↑	75.9	76.2	77.2	78.2	77.2	80.6	79.7	80.5	81.7	**82.4**	77.4
	Recall↑	89.1	91.1	90.1	89.1	87.8	**95.2**	91.1	93.1	90.9	94.5	87.5
	Decay↓	**0.0**	7.0	0.9	4.1	5.0	2.2	1.9	4.4	2.2	5.5	6.7
$\mathcal{F}$	Mean↑	72.1	70.6	74.5	75.9	77.4	75.5	77.4	79.5	80.5	**80.7**	77.3
	Recall↑	83.4	83.5	84.4	84.7	84.4	87.9	85.8	89.5	85.1	**90.2**	85.7
	Decay↓	1.3	7.9	**-0.2**	3.5	3.3	2.4	1.6	5.0	0.6	4.5	4.2

References

1. Chen, H., Li, Y.: Progressively complementarity-aware fusion network for rgb-d salient object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 3051–3060 (2018) 2, 3
2. Chen, H., Li, Y.: Three-stream attention-aware network for rgb-d salient object detection. IEEE Transactions on Image Processing **28**(6), 2825–2835 (2019) 2, 3
3. Chen, H., Li, Y., Su, D.: Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection. Pattern Recognition **86**, 376–385 (2019) 2, 3
4. Cheng, Y., Fu, H., Wei, X., Xiao, J., Cao, X.: Depth enhanced saliency detection method. In: International Conference on Internet Multimedia Computing and Service. p. 23 (2014) 2, 3
5. Cong, R., Lei, J., Zhang, C., Huang, Q., Cao, X., Hou, C.: Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion **23**(6), 819–823 (2016) 2, 3
6. Faisal, M., Akhter, I., Ali, M., Hartley, R.: Exploiting geometric constraints on dense trajectories for motion saliency. arXiv preprint arXiv:1909.13258 (2019) 3, 4
7. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: Proceedings of IEEE International Conference on Computer Vision. pp. 4548–4557 (2017) 2
8. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421 (2018) 2
9. Fan, D.P., Lin, Z., Zhao, J.X., Liu, Y., Zhang, Z., Hou, Q., Zhu, M., Cheng, M.M.: Rethinking rgb-d salient object detection: Models, datasets, and large-scale benchmarks. arXiv preprint arXiv:1907.06781 (2019) 2, 3
10. Han, J., Chen, H., Liu, N., Yan, C., Li, X.: Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. IEEE Transactions on Cybernetics **48**(11), 3171–3183 (2017) 2, 3
11. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: Proceedings of International Conference on Image Processing. pp. 1115–1119 (2014) 2, 3
12. Jun Koh, Y., Kim, C.S.: Primary object segmentation in videos based on region augmentation and reduction. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 3442–3450 (2017) 3, 4

13. Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F.: See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 3623–3632 (2019) 3, 4
14. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: Rgbd salient object detection: A benchmark and algorithms. In: Proceedings of European Conference on Computer Vision. pp. 92–109 (2014) 2, 3
15. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 724–732 (2016) 2
16. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection. In: Proceedings of IEEE International Conference on Computer Vision. pp. 7254–7263 (2019) 2, 3
17. Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., Yang, Q.: Rgbd salient object detection via deep fusion. IEEE Transactions on Image Processing **26**(5), 2274–2285 (2017) 2, 3
18. Siam, M., Jiang, C., Lu, S., Petrich, L., Gamal, M., Elhoseiny, M., Jagersand, M.: Video object segmentation using teacher-student adaptation in a human robot interaction (hri) setting. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 50–56. IEEE (2019) 3, 4
19. Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M.: Pyramid dilated deeper convlstm for video salient object detection. In: Proceedings of European Conference on Computer Vision. pp. 715–731 (2018) 3, 4
20. Sun, D., Yang, X., Liu, M., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. p. 89348943 (2018) 3
21. Tokmakov, P., Alahari, K., Schmid, C.: Learning video object segmentation with visual memory. In: Proceedings of IEEE International Conference on Computer Vision. pp. 4481–4490 (2017) 3, 4
22. Tokmakov, P., Schmid, C., Alahari, K.: Learning to segment moving objects. International Journal of Computer Vision **127**(3), 282–301 (2019) 3, 4
23. Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S.C., Ling, H.: Learning unsupervised video object segmentation through visual attention. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 3064–3074 (2019) 2, 3, 4
24. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018) 2
25. Yang, Z., Wang, Q., Bertinetto, L., Hu, W., Bai, S., Torr, P.H.: Anchor diffusion for unsupervised video object segmentation. In: Proceedings of IEEE International Conference on Computer Vision. pp. 931–940 (2019) 3, 4
26. Zhao, J.X., 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: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2019) 2, 3
27. Zhou, T., Wang, S., Zhou, Y., Yao, Y., Li, J., Shao, L.: Motion-attentive transition for zero-shot video object segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 2, p. 3 (2020) 2, 3, 4
28. Zhu, C., Li, G.: A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In: Proceedings of IEEE International Conference on Computer Vision. pp. 3008–3014 (2017) 2, 3

29. Zhu, C., Li, G., Wang, W., Wang, R.: An innovative salient object detection using center-dark channel prior. In: Proceedings of IEEE International Conference on Computer Vision. pp. 1509–1515 (2017) 2, 3