Weakly Supervised Video Salient Object Detection
(Supplementary Materials)

Wangbo Zhao[1] Jing Zhang[2,3] Long Li[1] Nick Barnes[2] Nian Liu[4] Junwei Han[1]✉ *
[1] The Brain and Artificial Intelligence Laboratory, Northwestern Polytechnical University
[2] Australian National University [3] CSIRO, Australia
[4] Inception Institute of Artificial Intelligence
{wangbo.zhao96, zjnwpu, longli.nwpu, liunian228, junweihan2010}@gmail.com,
nick.barnes@anu.edu.au

Abstract

In this supplementary material, we further show our weak label (the fixation guided scribble annotation) and configuration comparison of our method and existing methods. We also show other prediction boosting techniques and more qualitative comparison.

1. Fixation Guided Scribble Annotation

In Figure 1, we visualize more samples to show the scribble annotation in the proposed dataset. During annotating, an annotator is required to look the at RGB image(Figure 1 (a)) and the fixation map (Figure 1 (b)) at the same time. Based on this, the annotator annotates the foreground in objects with peak response regions and background in other regions, and we then obtain our scribble annotation (Figure 1 (d)). Note that, in the whole process of annotating, the clean GT (Figure 1 (c)) is not available to the annotator. Our dataset will be released to the public.

2. Configuration Comparison

We show more detailed configurations of our method and competing methods in Table 1, which clearly shows that our model has the cheapest condfiguration, leading to the least effort of annotation. Meanwhile, the efficiency of our model during inference further illustrates the effectiveness of our solution.

3. Effectiveness of boosting

Apart from the proposed saliency boosting strategy, we also try a traditional boosting strategy, named Ours-b. In this setting, we first use our scribble label to train the model. After several epochs of training, we adopt the prediction from the model as the pseudo label to finetune the whole model. From Table 2, we can find that Ours-b can bring improvements to on VOS [8], DAVIS [12] and FBMS [11]. While our boosting strategy can bring improvements and show effectiveness on more datasets. In Figure 2, we illustrate the comparison of pseudo labels from our strategy (Figure 2 (c)) and from the traditional strategy (Figure 2 (d)). It is clear that, our pseudo labels are closer to the clean GT (Figure 2 (b)).

4. Qualitative Comparison

In Figure 3, we visualize more samples from the testing sets of DAVSOD and DAVIS to compare our performance with competing techniques to further show the superior performance of our solution.

References

[1] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015. 3

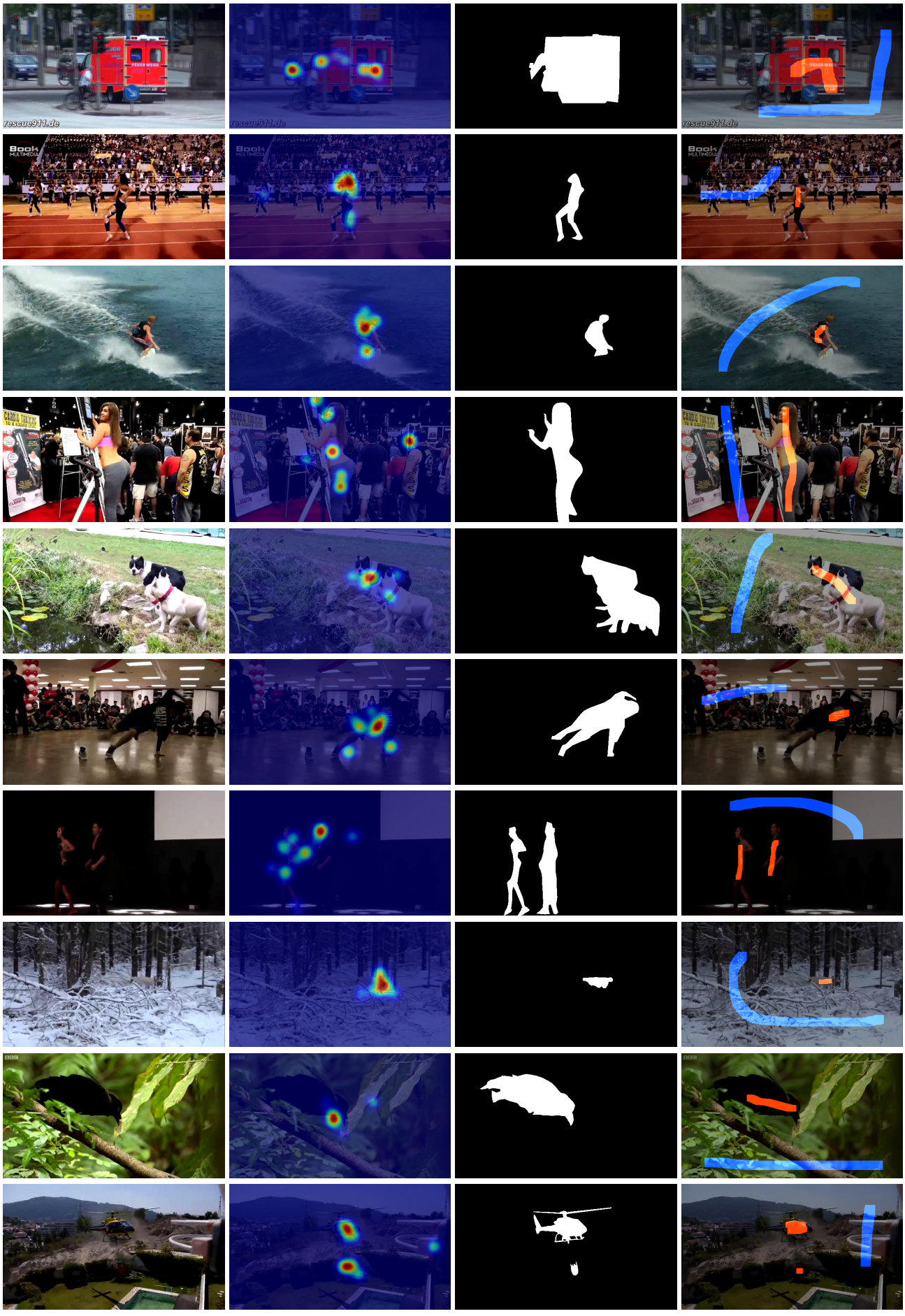
[2] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8554–8564, 2019. 3, 5

[3] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *AAAI*, pages 10869–10876, 2020. 3, 5

[4] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2192–2199, 2013. 3

[5] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient

*Corresponding author: Junwei Han *(junweihan2010@gmail.com)*



(a) Image (b) Fixation (c) Clean GT (d) Our GT

Table 1. Network setting of competing methods and ours. T.D. = Training data. DUTS [15]; MSRA10K [1]; S-DUTS [21]; HKU-IS [6]; DUTO [20]; DAVIS [12]; DAVSOD [2]; VOS [8]; FBMS [11]; ViSal [16]; SegV2 [4]

Method	**Ours**	**SSOD[21]**	**EGNet[22]**	**SCRN[18]**	**PoolNet[9]**	**FCNS[17]**
Time (s)	0.035	0.166	0.020	0.029	0.033	0.470
Backbone	ResNet-50	VGG-16	ResNet-50	ResNet-50	ResNet-50	VGG-16
T.D.	S-DUTS DAVSOD-S DAVIS-S	S-DUTS DAVSOD-S DAVIS-S	DUTS	DUTS	DUTS	MSRA10K DUTO FBMS SegV2
Method	**PDB[14]**	**FGRN[5]**	**MGA[7]**	**RCRNet[19]**	**SSAV[2]**	**PSCA[3]**
Time (s)	0.050	0.090	0.070	0.037	0.050	0.010
Backbone	ResNet-50	ResNet-101	ResNet-101	ResNet-50	ResNet-50	MobileNetV3
T.D.	pretrain:DUTO MSRA10K train:DAVIS	SegV2 FBMS DAVIS	DUTS DAVIS FBMS	MSRA10K HKU-IS train:VOS DAVIS FBMS	DUTO DAVSOD DAVIS	DUTS DAVSOD DAVSOD
Method	**TENet[13]**	**MuG[10]**				
Time (s)	0.060	0.600				
Backbone	ResNet-50	ResNet-50				
T.D.	DUTS DAVIS DAVSOD	OxUvA				

Table 2. Performance of our ablation study related experiments.

Method	VOS			DAVIS			DAVSOD			FBMS			SegV2			ViSal		
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
Ours	0.750	0.666	0.091	0.828	0.779	0.037	**0.705**	**0.605**	**0.103**	0.778	0.786	**0.072**	0.804	0.738	0.033	0.857	0.831	0.041
Ours-b	0.753	0.678	0.089	0.836	**0.800**	**0.036**	0.690	0.575	0.110	0.786	0.778	0.075	0.804	0.728	0.033	0.836	0.808	0.054
Ours*	**0.765**	**0.702**	**0.089**	**0.846**	0.793	0.038	0.694	0.593	0.115	**0.803**	**0.792**	0.073	**0.819**	**0.762**	**0.033**	**0.883**	**0.875**	**0.035**

object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3243–3252, 2018. 3

[6] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5455–5463, 2015. 3

[7] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7274–7283, 2019. 3, 5

[8] Jia Li, Changqun Xia, and Xiaowu Chen. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. *IEEE Transactions on Image Processing*, 27(1):349–364, 2017. 1, 3

[9] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3917–3926, 2019. 3

[10] Xiankai Lu, Wenguan Wang, Jianbing Shen, Yu-Wing Tai, David J Crandall, and Steven CH Hoi. Learning video object segmentation from unlabeled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8960–8970, 2020. 3

[11] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1187–1200, 2013. 1, 3

[12] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016. 1, 3

[13] Sucheng Ren, Chu Han, Xin Yang, Guoqiang Han, and Shengfeng He. Tenet: Triple excitation network for video salient object detection. *arXiv preprint arXiv:2007.09943*, 2020. 3, 5

[14] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 715–731, 2018. 3

[15] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan. Learning to detect salient objects with image-level supervision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3796–3805, 2017. 3

[16] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 24(11):4185–4196, 2015. 3, 5

[17] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1):38–49, 2017. 3

[18] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7264–7273, 2019. 3

[19] Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7284–7293, 2019. 3, 5

v (a) Image (b) GT (c) boosted (d) traditional

Figure 2. Boosting strategy comparison.

[20] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013. 3

[21] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12546–12555, 2020. 3, 5

[22] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8779–8788, 2019. 3

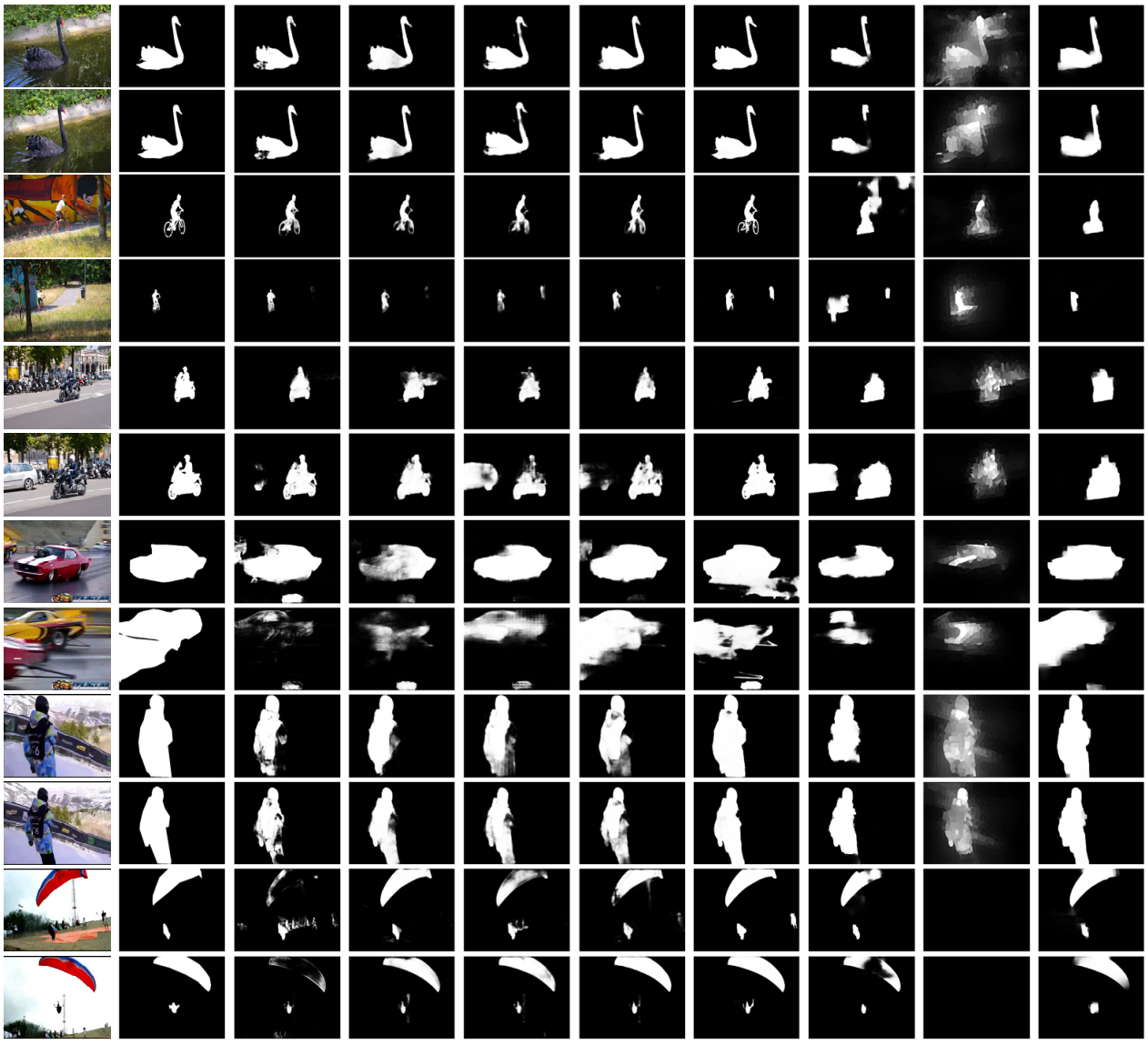


Figure 3. Qualitative comparison with state-of-the-art video salient object detection methods. MGA [7]; RCRNet [19]; SSAV [2]; PSCA [3]; TENet [13]; SSOD [21]; GF [16].