Supplementary Material
Semi-supervised Video Shadow Detection via Image-assisted Pseudo-label Generation

Anonymous Author(s)
Submission Id: 1353*

In this supplementary material, we first elaborate the details of the network architecture of our proposed STANet and MPLNet in Section 1. Then, in Section 2, we present more ablation experiments. Finally, more visualization results about video pseudo-label generation and more qualitative comparison results about video shadow detection are provided in Section 3.

1 DETAILS OF NETWORK ARCHITECTURE

STANet. As presented in Table 1, our STANet consists of three components: a shared feature extractor, a MARD module and a pixel-wise decoder. Our feature extractor is based on the ResNeXt-101, which also includes an additional convolution layer to align the number of channels and four bottleneck layers. The MARD module is designed with three consecutive convolutions and deformable convolution (DCN) layers to refine the offset and align features. In the experiments, we use the DCN with 8 deformable groups. In the decoder, we progressively upsample the aligned feature maps with the bilinear kernel to generate video pseudo-labels with the same size as the original inputs. Note that, we use a dropout layer to increase the generalization ability and generate the uncertainty map.

MPLNet. Similar with STANet, the architecture of our MPLNet is shown in Table 2. For simplification, we use the "MARD1" and "MARD2" to denote the MARD Module in Table 1.

2 MORE ABLATION EXPERIMENTS

we perform the following ablation experiments to explore the settings of the hyperparameter n in STANet and the iteration number in MARD module in this paper.

Hyperparameter n. We conduct experiments on pseudo-label generation with the hyperparameter $n = \{1, 2, 3\}$ to analyze the changes of performance, time-consumption and memory-consumption. The results are summarized in Table 3. The table shows that, the larger n will lead to the better performance while result in the greater resource consumption, and $n = 2$ is a good trade-off between effectiveness and efficiency.

The iteration number. The Table 4 shows the results of different number of iterations in the MARD module on video shadow detection. We can see 3 times of iteration achieves the best performance.

In addition, we also perform experiments with different optical flow estimation methods: GMA [4], FlowNet2 [3], and PWC-Net [6] for proving the insensitivity of our MARD module for optical flow estimation. From the results in Table 5 we can observe that, there is no significant difference between the performance using the different optical flow estimation network, which demonstrates that the performance of our MARD module dose not heavily rely on the optical flow.

Table 1: The architecture of our STANet.

	Layer Name	Output Size	Operation
Feature extractor	Conv	400 × 400 × 3	Conv(1 × 1)
	Backbone	-	ResNeXt-101
	Bottle1	100 × 100 × 64	Conv(3 × 3)
	Bottle2	50 × 50 × 64	Conv(3 × 3)
	Bottle3	25 × 25 × 64	Conv(3 × 3)
	Bottle4	25 × 25 × 64	Conv(3 × 3)
MARD Module	Off1	50 × 50 × 144	Conv(3 × 3)
	DCN1	50 × 50 × 64	DCN(3 × 3)
	Off2	50 × 50 × 144	Conv(3 × 3)
	DCN2	50 × 50 × 64	DCN(3 × 3)
	Off3	50 × 50 × 144	Conv(3 × 3)
	DCN3	50 × 50 × 64	DCN(3 × 3)
Decoder	Res1	25 × 25 × 64	Resblock(3 × 3)
	Up1	50 × 50 × 64	Upsample
	Res2	50 × 50 × 64	Resblock(3 × 3)
	Up2	100 × 100 × 64	Upsample
	Res3	100 × 100 × 64	Resblock(3 × 3)
	Up3	200 × 200 × 64	Upsample
	Conv4	200 × 200 × 64	Conv(3 × 3)
	Up4	400 × 400 × 64	Upsample
	Conv5	400 × 400 × 64	Conv(3 × 3)
	Drop	400 × 400 × 64	Dropout(0.1)
	Pred	400 × 400 × 3	Conv(3 × 3)

3 MORE VISUALIZATION RESULTS

In this section, we first provide more visualization results for our pseudo-label generation and uncertainty map estimation in Figure 1. Then, the more qualitative comparison results with FSDNet [2], BDRAR [10], DSD [9], TVSD-Net [1], RCRNet [8], MoNet [7], and COSNet [5] are presented in Figure 2.

REFERENCES

[1] Zhihao Chen, Liang Wan, Lei Zhu, Jia Shen, Huazhu Fu, Wennan Liu, and Jing Qin. 2021. Triple-cooperative video shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2715–2724.

[2] Xiaowei Hu, Tianyu Wang, Chi-Wing Fu, Yitong Jiang, Qiong Wang, and Pheng-Ann Heng. 2021. Revisiting shadow detection: A new benchmark dataset for complex world. *IEEE Transactions on Image Processing* 30 (2021), 1925–1934.

[3] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2462–2470.

[4] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. 2021. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9772–9781.

[5] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. 2019. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *Proceedings of the IEEE/CVF Conference*

Table 2: The architecture of our MPLNet.

	Layer Name	Output Size	Operation
Feature extractor	Backbone	-	ResNeXt-101
	Bottle1	100 × 100 × 64	Conv(3 × 3)
	Bottle2	50 × 50 × 64	Conv(3 × 3)
	Bottle3	25 × 25 × 64	Conv(3 × 3)
	Bottle4	25 × 25 × 64	Conv(3 × 3)
MARD1	-	25 × 25 × 64	-
MARD2	-	50 × 50 × 64	-
Mempry Propagation	DCN	25 × 25 × 256	DCN(3 × 3)
	Conv	25 × 25 × 1	Convblock(3 × 3)
Decoder	Res1	25 × 25 × 64	Resblock(3 × 3)
	Up1	50 × 50 × 64	Upsample
	Res2	50 × 50 × 64	Resblock(3 × 3)
	Up2	100 × 100 × 64	Upsample
	Res3	100 × 100 × 64	Resblock(3 × 3)
	Up3	200 × 200 × 64	Upsample
	Conv4	200 × 200 × 64	Conv(3 × 3)
	Up4	400 × 400 × 64	Upsample
	Conv5	400 × 400 × 64	Conv(3 × 3)
	Drop	400 × 400 × 64	Dropout(0.1)
	Pred1	25 × 25 × 3	Conv(3 × 3)
	Pred2	50 × 50 × 3	Conv(3 × 3)
	Pred3	100 × 100 × 3	Conv(3 × 3)
	Pred4	200 × 200 × 3	Conv(3 × 3)
	Pred5	400 × 400 × 3	Conv(3 × 3)
	Pred-final	400 × 400 × 3	Conv(1 × 1)

Table 3: Quantitative ablation results on ViSha about hyperparameter n. Here *Time* and *Memory* denote time-consumption and memory-consumption respectively, ↓ denotes the smaller is better.

n	*BER* ↓	*Time* ↓	*Memory* ↓
1	11.06	**2.15**	**3577M**
2	10.34	2.74	5014M
3	**10.29**	3.28	7635M

Table 4: Quantitative ablation results on ViSha about iteration number.

Iteration #	*BER* ↓
1	11.24
2	10.83
3	**10.34**
4	10.37

on Computer Vision and Pattern Recognition. 3623–3632.

[6] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8934–8943.

[7] Huaxin Xiao, Jiashi Feng, Guosheng Lin, Yu Liu, and Maojun Zhang. 2018. Monet: Deep motion exploitation for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1140–1148.

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

Table 5: Quantitative ablation results on ViSha about optical flow estimation method.

Method	*BER* ↓
GMA [4]	**10.34**
Flownet2 [3]	10.48
PWC-Net [6]	10.39

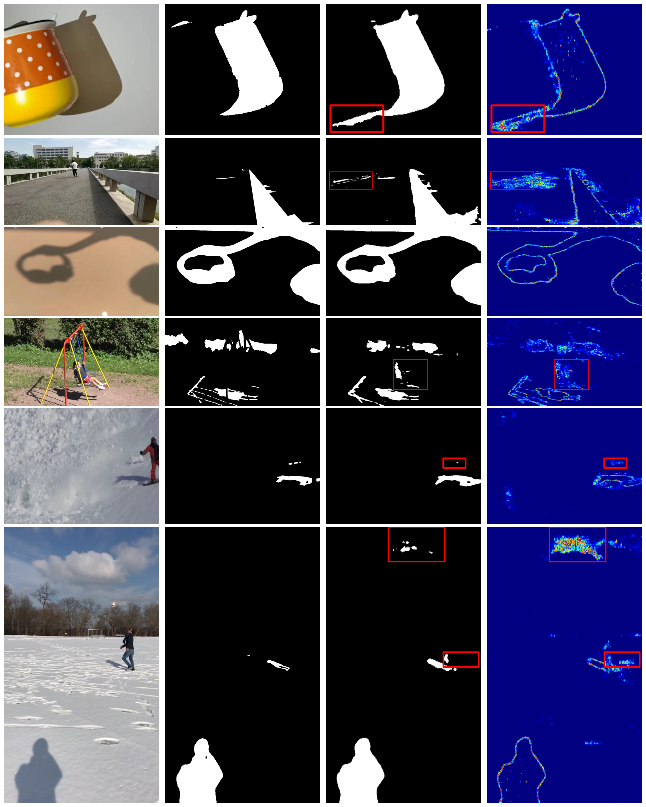

Figure 1: Visualization results of our pseudo-label generation and uncertainty map estimation. From left to right are: input video frame, the corresponding ground-truth, the generated pseudo-labels and the estimated uncertainty map.

[9] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson W.H. Lau. 2019. Distraction-Aware Shadow Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[10] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. 2018. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 121–136.

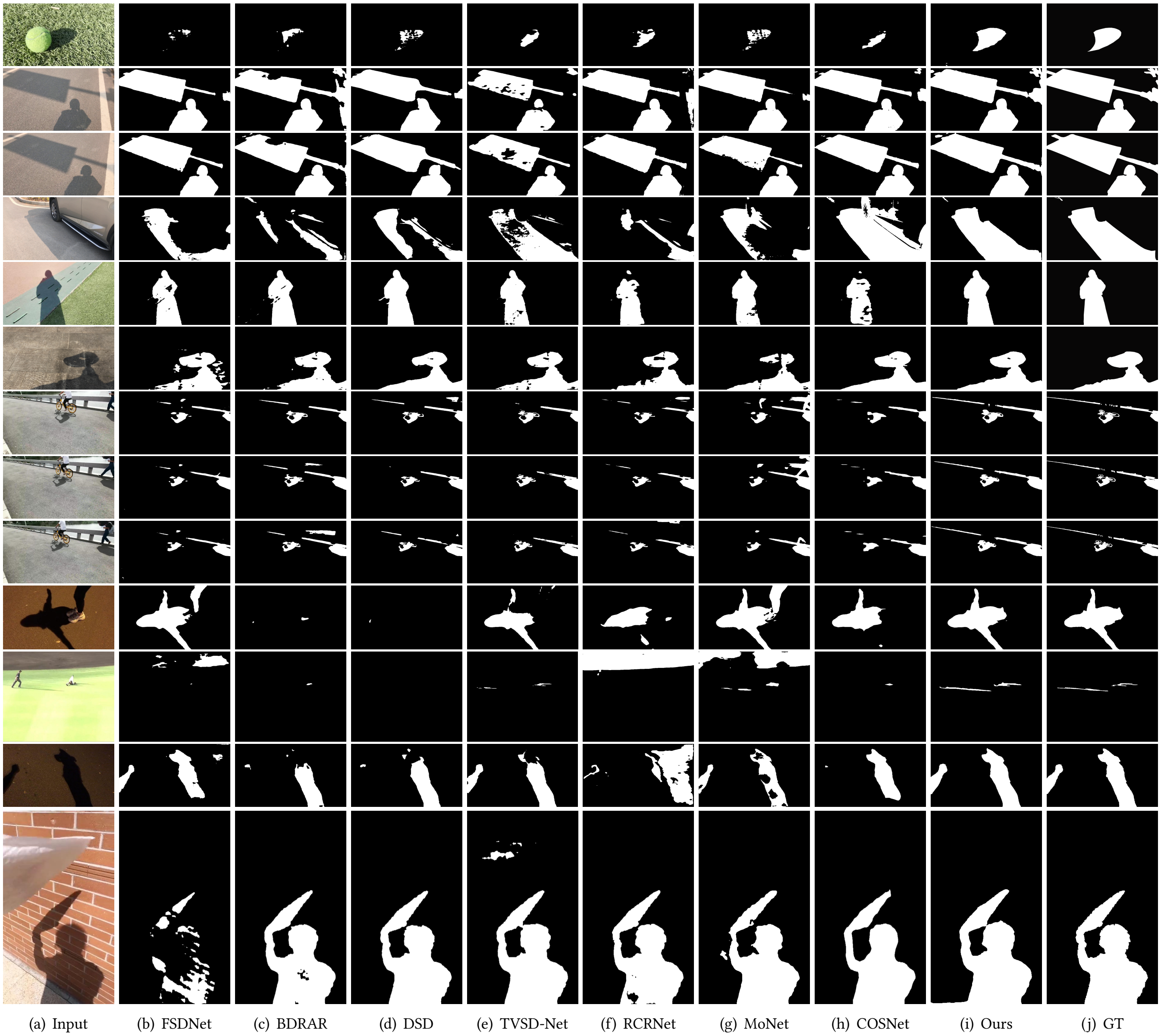

Figure 2: Qualitative comparison of shadow masks predicted by our method(i) and other SOTA methods (b-h) on ViSha [1] and RVSD datasets.