Learning Shadow Correspondence for Video Shadow Detection
Supplementary Material

Xinpeng Ding[1 [0000−0001−7653−1199]], Jingwen Yang[1 [0000−0003−2420−0130]], Xiaowei Hu[2 [0000−0002−5708−7018]], and Xiaomeng Li[1,3 [0000−0003−1105−8083]⋆]

[1] The Hong Kong University of Science and Technology
{xdingaf,jyangbv}@connect.ust.hk
[2] Shanghai AI Laboratory
huxiaowei@pjlab.org.cn
[3] The Hong Kong University of Science and Technology Shenzhen Research Institute
eexmli@ust.hk

This supplementary file provides more training details for our proposed brightness-invariant shadow-consistent correspondence method (BS-Cor). Then, we show the ablation study on grid size L and hyper-parameter λ and β. We also compare different generate optical flows. Finally, we show more visualization of video shadow detection results and some failure cases.

1. Training Details

Following the previous works [1, 5], we initialize the backbone with weights pre-trained on ImageNet [2]. In each training iteration, the input images are resized to a resolution of 416×416 for TVSD [1] and 320×320 for DSD [5]. Following [1, 5], we randomly apply horizontal flipping for samples in the training phase. The model is built with Pytorch [4] and is trained by two NVIDIA 3090 GPUs.

2. Ablation Study

Grid size L. Table 1 reports the results of using different grid sizes L (see Eq. 4 in the paper). The results show that large grid size improves the performance, where more correlations of pixels are extracted. We set L as 17×17 as the default setting in other experiments, since the performance gain is limited when L goes beyond 17×17.

Hyper-parameter λ and β. We compare the average score of IoU and TS of different λ in Eq. 9 and β in Eq. 7 in Fig. 1. $\lambda = 0$ provides a baseline model that does not adopt the SC-Cor loss. After we increase the weight of SC-Cor learning objective, we can clearly improve the performance, showing the effectiveness of the proposed SC-Cor. Finally, we achieve the best results by setting λ as ten and we use this value in other experiments. With the β increasing, the performance of video shadow detection first increases and then becomes stable. As shown in Table 1 (b), our method can achieve the best performance when β is 0.5.

⋆ Corresponding Authors

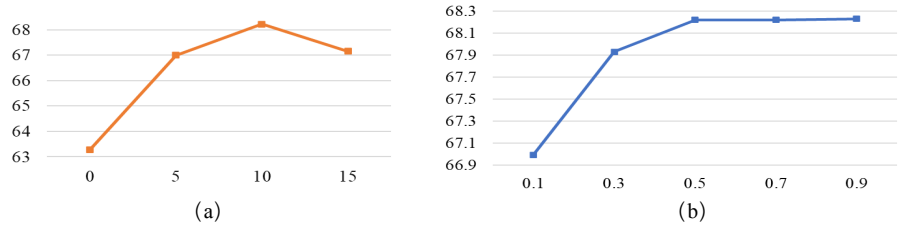

Fig. 1: The average score of IoU and TS of different (a) λ defined in Eq. 9 and (b) β defined in Eq. 7.

Table 1: **Ablation on grid size** L. Larger grid size brings higher performance. We set the grid size as 17×17 in other experiments.

	Frame-level		*Temporal-level*	
L	BER ↓	IoU [%] ↑	TS [%] ↑	AVG ↑
13×13	17.45	56.98	77.27	67.13
15×15	16.24	57.26	77.47	67.37
17×17	14.89	58.40	78.03	68.22
19×19	14.79	58.33	78.12	68.24

3. Visualization Results

Fig. 2 shows the visual comparison of video shadow detection results produced by different methods. Notably, our method can predict more consistent and accurate detection results than other existing methods.

4. Comparison of optical flows

Since the motions of shadows are hard to be captured on the RGB frames (Line387), we use consecutive GT labels. The optical flows generated by RGB are focus on objects, which can not capture shadows; see Fig. 4. We also use optical flow computed by RGB frames to compute TS: DSD = 62.3% TVSD = 51.5%. This result is not reasonable, since TVSD is the SOTA for video shadow detection while DSD is for image. Will add the comparison in paper.

5. Failure Cases

We present some failure cases in Fig. 3. Our method may misclassify dark objects as shadow regions when dark objects occupy the whole image. However, compared with other state-of-the-art methods, our method can achieve better performance.

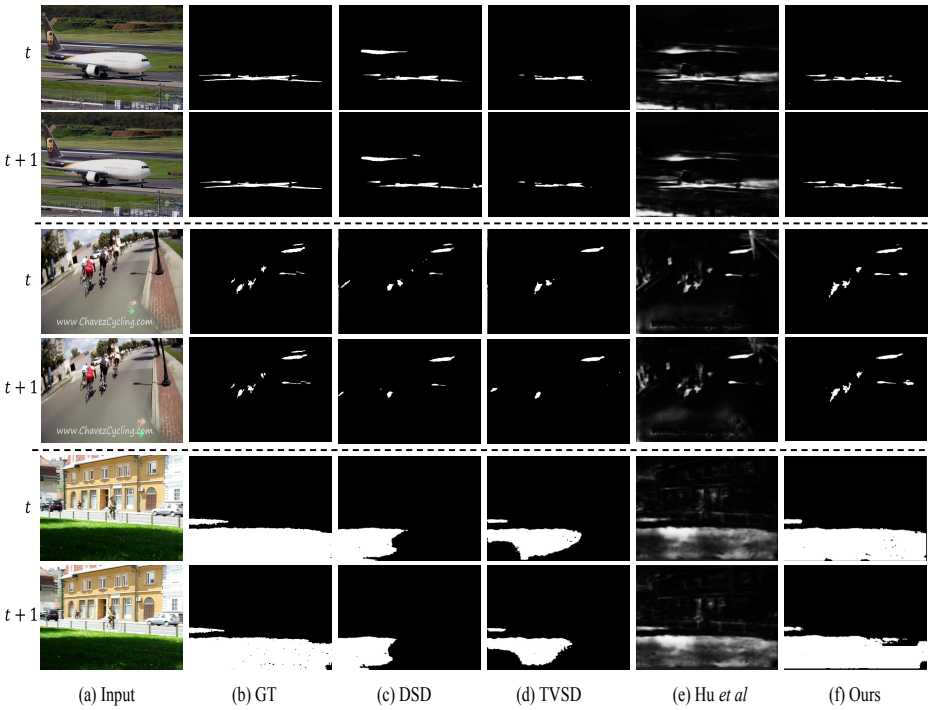

Fig. 2: Visual comparison of video shadow detection results produced by different methods. (a) is the input images and (b) is the ground-truth (GT) images. (c)-(f) are the results predicted by DSD [5], TVSD [1], Hu *et al.* [3], and our method, respectively. Our method takes the DSD as the basic network.

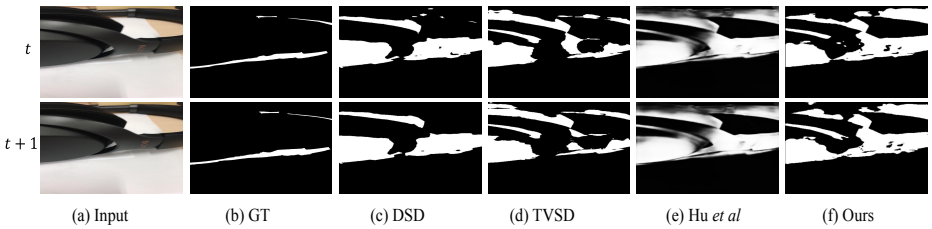

Fig. 3: Failure cases predicted by different models. (a) is the input images and (b) is the ground-truth (GT) images. (c)-(f) are the results predicted by DSD [5], TVSD [1], Hu *et al.* [3], and our method, respectively. Our method takes the DSD as the basic network.

Fig. 4: Comparison of the optical flows generated by RGB and ground-truth (GT) labels.

References

1. Chen, Z., Wan, L., Zhu, L., Shen, J., Fu, H., Liu, W., Qin, J.: Triple-cooperative video shadow detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2715–2724 (2021)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
3. Hu, S., Le, H., Samaras, D.: Temporal feature warping for video shadow detection. arXiv preprint arXiv:2107.14287 (2021)
4. Paszke, A., Gross, S., Chintala, S.: Pytorch deep learning framework. Web page (2017), http://pytorch.org/
5. Zheng, Q., Qiao, X., Cao, Y., Lau, R.W.: Distraction-aware shadow detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5167–5176 (2019)