—-Supplementary Material—-
Video Shadow Detection via Spatio-Temporal Interpolation Consistency Training

Xiao Lu[1], Yihong Cao[1], Sheng Liu[1], Chengjiang Long[2],
Zipei Chen[3], Xuanyu Zhou[1], Yimin Yang[4,5], Chunxia Xiao[3*]
[1]College of Engineering and Design, Hunan Normal University, Changsha, China
[2]Meta Reality Labs, Burlingame, CA, USA
[3]School of Computer Science, Wuhan University, Wuhan, Hubei, China
[4]Department of Computer Science, Lakehead University, Canada
[5]Vector Institute for Artificial Intelligence, Canada
{luxiao,caoyihong,liusheng,zhouxy}@hunnu.edu.cn, clong1@fb.com,
{czpp19,cxxiao}@whu.edu.cn, yyang48@lakeheadu.ca

Abstract

In this supplementary material, we firstly provide the ablation results on ViSha [1] in Section 1. Then we test the performance of our Spatial ICT and a semi-supervised semantic segmentation method CCT [2] for comparison in Section 2. In addition, We present the sensitivity analysis on the parameters of our method in Section 3, and provide some details on our VISAD dataset in Section 4.

1. Ablation study on ViSha

We also conduct the ablation study on ViSha to understand the behavior and effectiveness of each module we proposed.

Ablation study on SANet. The three modules, EDR, FFM and DAM are hierarchically added on the basic encoder-decoder network with a simple feature fusing structure as the ablation study on DS. The upper part of Table 1 summarizes the quantitative results, and the qualitative results are presented Fig.1.

From the results, we can see that all the three modules are very effective for promoting the performance, which demonstrates that they are necessary for our SANet for learning accurate shadow features. The visualization results presented in Fig.1 also verify the effectiveness of each module on the detection of details and small scale shadow regions.

Ablation study on the three consistency constraints. The three consistency constraints, the scale consistency constraint (SC), the spatial interpolation consistency constraint (SIC), and the temporal consistency constraint (TIC) are hierarchically added on the basic SANet. The lower part of Table 1 summarizes the quantitative results, and the qualitative results are presented Fig.2. The quantitative and qualitative results all demonstrate that the three consistency constraints are effective for boosting the performance of shadow detection.

ED	FFM	R	DAM	MAE↓	F_β ↑	IoU ↑	BER↓
✓				0.050	0.664	0.522	17.27
✓	✓			0.039	0.717	0.567	16.33
✓	✓	✓		0.037	0.742	0.583	14.34
✓	✓	✓	✓	**0.036**	**0.752**	**0.596**	**13.26**
	SC	TIC	SIC	MAE ↓	F_β ↑	IoU ↑	BER ↓
B				0.062	0.542	0.463	18.11
	✓			0.059	0.590	0.493	16.82
	✓	✓		0.052	0.593	0.490	16.76
	✓	✓	✓	**0.046**	**0.702**	**0.545**	**16.60**

Table 1. The upper part: results of ablation analysis on SANet pretrained on SBU and fine-tuned on ViSha, R: Refiner. The lower part: results of ablation analysis on STICT, B: basic SANet trained on SBU and tested on ViSha without fine-tuning.

2. Comparison of Spatial ICT with CCT [2]

To demonstrate the effectiveness of Spatial ICT for better generalization, we compare Spatial ICT with a semi-supervised semantic segmentation method Cross-Consistency Training (CCT) [2], where the coss-consistency regularization is enforced to encourage an invariant of the predictions over different perturbations applied to the outputs of the encoder. We use the training set in SBU [3] and the training set in ViSha as the labeled and

*Corresponding author.

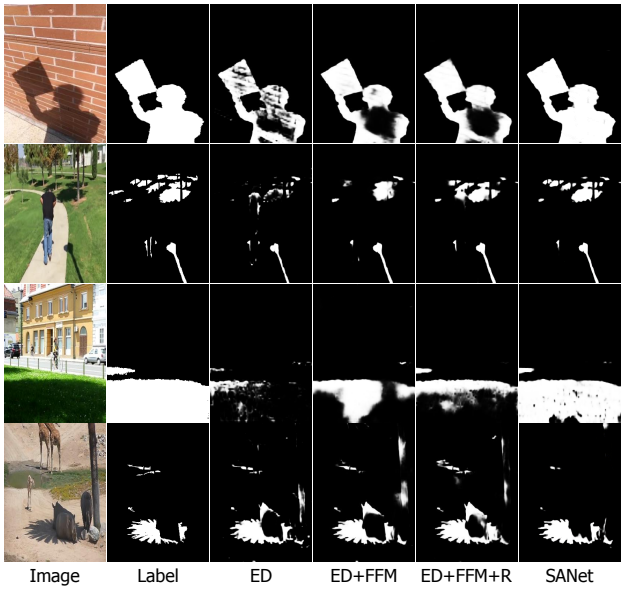

Figure 1. Visualization results of ablation study on SANet.

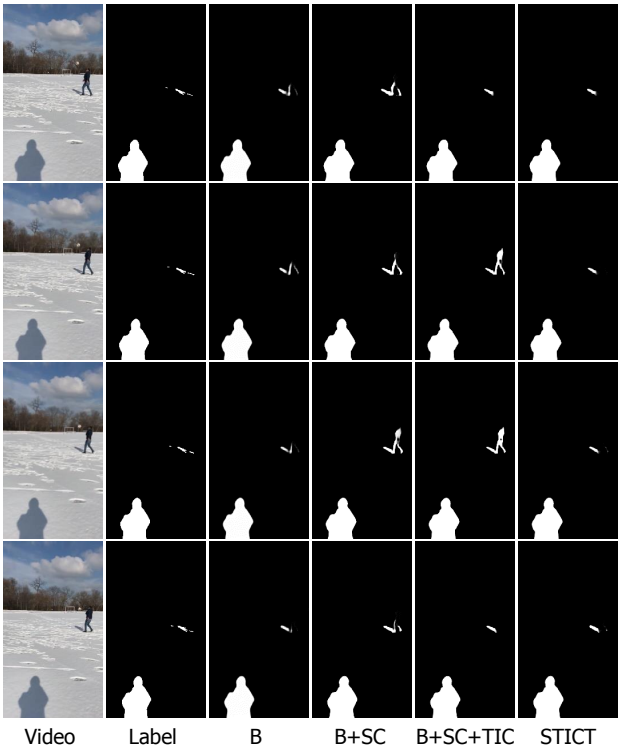

Figure 2. Visualization results of ablation study on the three consistency constraints.

unlabeled dataset for training, and test the models on the testing set in ViSha. In our Spatial ICT, the spatial interpolation is conducted on the outputs of the encoder, and only the spatial interpolation consistency constraint is used as the unsupervised loss for updating the student network. The test results are presented in Table 2, from which we can see that our Spatial ICT is more effective than CCT in our shadow region segmentation task.

	MAE↓	F_β ↑	IoU ↑	BER ↓
CCT [2]	0.098	0.494	0.331	20.61
Spatial ICT	0.052	0.612	0.502	16.24

Table 2. Performance of our Spatial ICT vs. CCT.

3. Sensitivity analysis on the parameters

All the experiments are conducted on DS for the ablation study on the parameters.

3.1. Sensitivity analysis on the spatial interpolation parameter d

As we plug the spatial interpolation module in the bottleneck between encoder and decoder, the width and height of feature map $\mathbf{F}$ are both 11, we test the sensitivity of our method's performance on different values of parameter d ($d \in [3, 5, 7]$). The test results are shown in Table 3. In theory, the spatial interpolation is more effective when d is larger, as the found unrelated point for interpolation in a larger neighboring area would be more likely to be in different class. However, from the results presented in Table 3, the best performance is achieved when $d = 3$. The reason is that our feature map is relatively small, and it needs a padding operation to calculate the spatial correlation when d gets larger, and then most of the points are interpolated with a zero point, which results in the failure of spatial interpolation.

	MAE↓	F_β ↑	IoU ↑	BER↓
$d = 3$	**0.065**	0.646	**0.370**	**14.17**
$d = 5$	0.071	**0.647**	0.366	15.89
$d = 7$	0.078	0.592	0.346	16.12

Table 3. Performance of our method vs. different values of d in the spatial interpolation module, the best results are highlighted with bold.

3.2. Sensitivity analysis on temporal interpolation parameter k

We test the sensitivity of our method's performance to different values of parameter k ($k \in [1, 2, 3, 4, 5]$) in Eq.(8). Since the images are sampled from the videos with a sampling rate $1/8$, $k = n$ means that we use a frame and its forward $8n^{th}$ frame and its backward $8n^{th}$ frame as the consecutive three frames for computing the temporal interpolation consistency loss. The test results are shown in Table

4. It can be observed that the performance of our proposed method decreases gradually with k increasing. The reason is that the larger motion between adjacent frames leads to more inaccurate optical flow, which makes the temporal interpolation consistency constraint difficult to maintain.

3.3. Sensitivity analysis on the weight parameters β_{max} and t_{max} in the Gaussian ramp up function for computing $\beta(t)$

We test the sensitivity of our method's performance to different values of parameters β_{max} and t_{max} in the Gaussian ramp-up function for updating the consistency loss weight $\beta(t) = \beta_{max} e^{-5(1-t/t_{max})^2}$, for analyzing the importance of the consistency loss in different training time period. The upper part of Table 5 presents the results of different values of β_{max} when $t_{max} = 10$, and the lower part of 5 presents the results of different values of t_{max} when $\beta_{max} = 1$, from which we can choose $\beta_{max} = 1$ and $t_{max} = 10$ for the best trade-off between the four metrics.

3.4. Sensitivity analysis on the weight parameters η_1, η_2 and η_3

We also test the sensitivity of our method's performance to different values of the weight parameters η_1, η_2, and η_3 for each consistency loss in Eq.(1). The results are presented in Table 6. It can be observed that we can choose the following three weight parameters, $\eta_1 = \eta_2 = \eta_3 = 1$, for the best trade-off between the four metrics.

3.5. Sensitivity analysis on the decay parameters η in EMA for updating the teacher network

We test the sensitivity of our method's performance to different values of the decay parameter η in EMA. The experimental results tested on different values of η are presented in Table 7. We choose $\eta = 0.999$ for the best trade-off between the four metrics.

4. More Details about the VISAD Dataset

Processing the Bonnet Region in DS. Considering that the front of the bonnet in the image, which is usually misrecognized as a shadow region, is settled in a video, we labeled the bonnet region in each annotated frame to mask the bonnet region during prediction phase for calculating the metric.

Dataset Analysis. To validate the diversities and challenges of our VISAD dataset, we analyze the shadow regions by using the connected component analysis technology, and we show the statistics as follows:

Shadow Scale. We define the scale of a shadow region by:

$$s = max\left(\frac{h_{bbox}}{H}, \frac{w_{bbox}}{W}\right), \quad (1)$$

where (h_{bbox}, w_{bbox}) and (H, W) are the height and width of the minimum enclosing rectangle of a shadow region and that of the image, respectively. The statistics of scale distribution for DS and MOS are shown in Fig.3(a) and (d), respectively. We can see that there are various scales of

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
MAE↓	**0.065**	0.079	0.082	0.089	0.117
F_β ↑	**0.646**	0.571	0.522	0.483	0.341
IoU↑	**0.370**	0.342	0.306	0.330	0.219
BER ↓	**14.17**	16.12	19.28	22.84	23.05

Table 4. Performance of our method vs. different values of k in Eq.(8), the best results are highlighted with bold.

$\beta_{max} =$	0.1	0.5	1	2	3
MAE↓	0.068	0.073	0.065	**0.064**	0.072
F_β ↑	0.597	0.621	**0.646**	0.436	0.520
IoU↑	0.373	**0.375**	0.370	0.335	0.374
BER↓	15.95	17.84	**14.17**	15.21	15.01
$t_{max} =$	5	10	15	20	30
MAE↓	0.072	**0.065**	0.068	0.068	0.081
F_β ↑	0.584	**0.646**	0.635	0.622	0.573
IoU↑	0.363	0.370	**0.373**	0.371	0.366
BER↓	15.26	**14.17**	14.22	14.98	17.65

Table 5. The upper part: performance of our method vs. different values of β_{max}. The lower part: performance of our method vs. different values of t_{max}.

		0.01	0.1	0.5	1	2
$\eta_1 = 0, \eta_2 = 0, \eta_3 =$	MAE↓	0.094	0.093	0.093	**0.092**	0.093
	F_β ↑	0.504	0.510	**0.519**	0.518	0.511
	IoU ↑	0.306	0.304	0.310	**0.311**	0.308
	BER↓	17.39	17.01	17.03	**16.78**	17.25
$\eta_1 = 0, \eta_3 = 1, \eta_2 =$	MAE↓	0.089	0.084	**0.078**	0.079	0.082
	F_β ↑	0.519	0.543	0.569	0.587	**0.590**
	IoU↑	0.309	0.312	0.318	**0.320**	0.313
	BER↓	16.57	16.77	17.02	**16.29**	17.22
$\eta_2 = \eta_3 = 1, \eta_1 =$	MAE↓	0.081	0.068	0.067	**0.065**	0.070
	F_β ↑	0.582	0.631	**0.649**	0.646	0.633
	IoU↑	0.329	0.364	0.361	0.370	**0.372**
	BER↓	16.75	15.86	16.23	**14.17**	14.39

Table 6. Performance of our method vs. different values of η_1, η_2 and η_3 in Eq.(1), the best results are highlighted with bold.

	$\eta = 0.99$	$\eta = 0.999$	$\eta = 0.9999$
MAE↓	0.062	**0.065**	0.072
F_β ↑	0.635	**0.646**	0.632
IoU↑	**0.373**	0.370	0.369
BER↓	15.29	**14.17**	14.30

Table 7. Performance of our method vs. different values of the decay parameter η in EMA.

shadow regions in DS and MOS, and most of them are small scale regions.

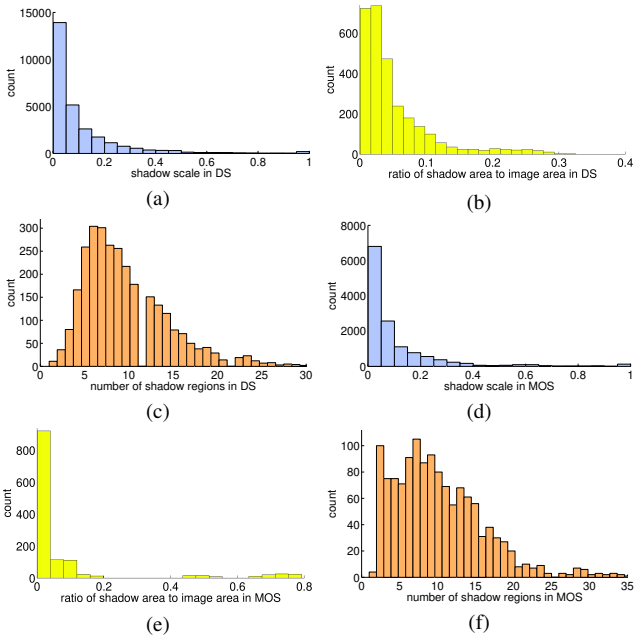

Figure 3. Statistics of our VISAD dataset. The scale distribution in DS (a) and in MOS (d). The shadow area distribution in DS (b) and in MOS (e), and the number of shadow regions distribution in DS (c) and in MOS (f).

Shadow Area. We define the area of the shadow region as a proportion of shadow pixels in the image. In Fig.3(b) and (e), we can see that the shadows in DS are mainly in small areas, in the range of $(0, 0.3]$, while that in MOS vary in a wide range with the majority falling in the range of $(0, 0.4]$. Such small shadow regions can be easily cluttered with diverse background objects/scenes.

Number of Shadow Regions. We define the number of shadow regions as the total number of connected components in an image. In Fig.3(c) and (f), we can see that there are more than five shadow regions in most of images, and even thirty in some of the images. The large quantity of small area shadow may degrade the performance of the shadow detection algorithm.

In summary, the shadows with scale variance, small area and large quantity in our VISAD dataset are the main challenges that may affect the performance of the algorithm.

References

[1] Zhihao Chen, Liang Wan, Lei Zhu, Jia Shen, Huazhu Fu, Wennan Liu, and Jin Qin. Triple-cooperative video shadow detection. In *2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[2] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 1, 2

[3] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2018. 1