Supplemental Material for "Is It Necessary to Transfer Temporal Knowledge for Domain Adaptive Video Semantic Segmentation?"

Xinyi Wu[1], Zhenyao Wu[1], Jin Wan[2], Lili Ju[1], and Song Wang[1]

[1] University of South Carolina
{xinyiw,zhenyao}@email.sc.edu ju@math.sc.edu songwang@cec.sc.edu
[2] Beijing Jiaotong University
jinwan@bjtu.edu.cn

In this supplementary material, we provide the following additional results and discussions of this paper.

1 Additional Results

1.1 Quantitative results

The impact of optical-flow computation for testing. In the main paper, our reported results are based on per-frame inference without optical-flow computation. Here, we further study the impact of optical-flow computation for testing. The comparison results under the VIPER → Cityscapes scenario and the SYNTHIA → Cityscapes scenario are shown in Table S1 and Table S2, respectively. To be specific, the "I2VDA (two-frame)" is implemented by a non-parametric fusion defined as:

$$P = \mathcal{M}(I_f) + \gamma \mathcal{M}(\mathcal{W}(I_{f-1}, F)), \tag{1}$$

where $\mathcal{W}(\cdot,\cdot)$ is the warping function and $\mathcal{M}$ is the segmentation network that have been defined in the main paper, I_f is the current frame, I_{f-1} represents its previous frame, F is the optical flow between I_{f-1} and I_f, and γ is set to 0.5 to balance the fusion. We observe that the improvements for both scenarios are not very obvious by further using two frames and computing optical flow for testing. These results, to some extent, indicate that our proposed temporal augmentation strategy is effective to help learn diverse temporal patterns during training thus there is no need to explicitly consider the temporal consistency during testing.

1.2 Qualitative results

Static results. For qualitative comparisons, we first provide some visualization results of independent frames for both VIPER → Cityscapes and SYNTHIA → Cityscapes scenarios scenarios in Figure S1 and Figure S2, respectively. It is observed that our method achieves more accurate results, especially on the small-size classes such as person and traffic light.

Table S1. Ablation study on the impact of optical-flow computation for testing under the VIPER → Cityscapes scenario.

Methods	road	sidewalk	building	fence	traffic light	traffic sign	vegetation	terrain	sky	person	car	truck	bus	motorcycle	bicycle	**mIoU (%)**
single	84.78	36.09	84.02	28.02	36.46	36.02	85.89	48	73.97	63.18	81.87	33.0	51.75	39.94	0.17	51.18
two-frame	85.13	36.66	84.11	26.36	36.18	35.88	86.10	33.13	74.29	63.33	82.12	32.80	52.38	39.42	0.22	51.21

Table S2. Ablation study on the impact of optical-flow computation for testing under the SYNTHIA → Cityscapes scenario.

Methods	road	sidewalk	building	pole	traffic light	traffic sign	vegetation	sky	person	rider	car	**mIoU (%)**
single	89.88	40.54	77.58	27.26	18.69	23.60	76.07	76.34	48.48	22.39	82.13	53.00
two-frame	89.95	40.74	77.58	27.53	17.92	23.12	76.22	76.60	48.47	22.42	82.20	53.00

Dynamic results. As a video-level task, we further provide dynamic visualization with 10 video samples from Cityscapes in "CVPR2022_735_video.mp4". The dynamic visualization shows that our method achieves more smooth and accurate results compared with existing video-to-video based method DA-VSN [2]. Note that, for clarity, we pause each video sample for 1 second at the frame with the ground truth.

2 Limitation and discussion

An accurate pretrained optical flow estimator (*e.g.*, FlowNet [3]) is the cornerstone of our method. The estimated flow is employed to synthesize the intermediate frames for temporal augmentation and to warp the prediction for ensuring target temporal consistency as well. Since the flow estimator is trained on additional synthetic datasets such as Sintel [1], employing the trained model weights in the target domain for optical flow estimation also suffers from a domain gap. An unsatisfactory optical flow estimation will mislead the training of the domain adaptive semantic segmentation which is a limitation of our method. Pretraining the flow estimator directly on the target domain (*e.g.*, Cityscapes) in an unsupervised manner might be a good future direction to bridge the domain gap caused by optical flow estimation.

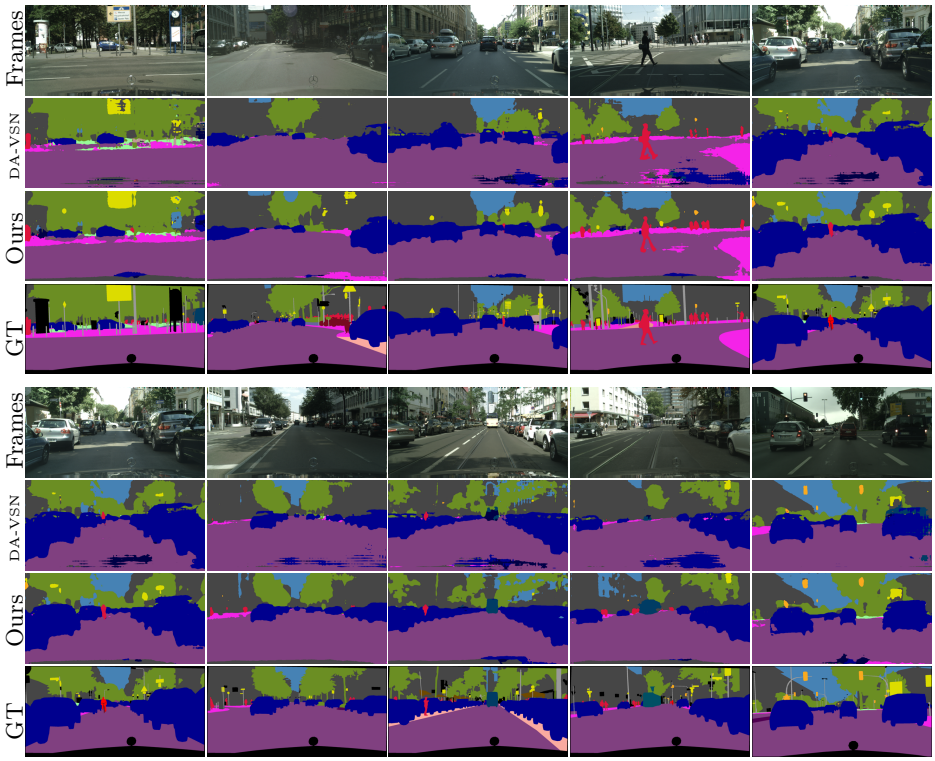

Fig. S1. Qualitative comparison results on the VIPER → Cityscapes domain adaptive video segmentation task using 10 independent samples from the Cityscapes validation set.

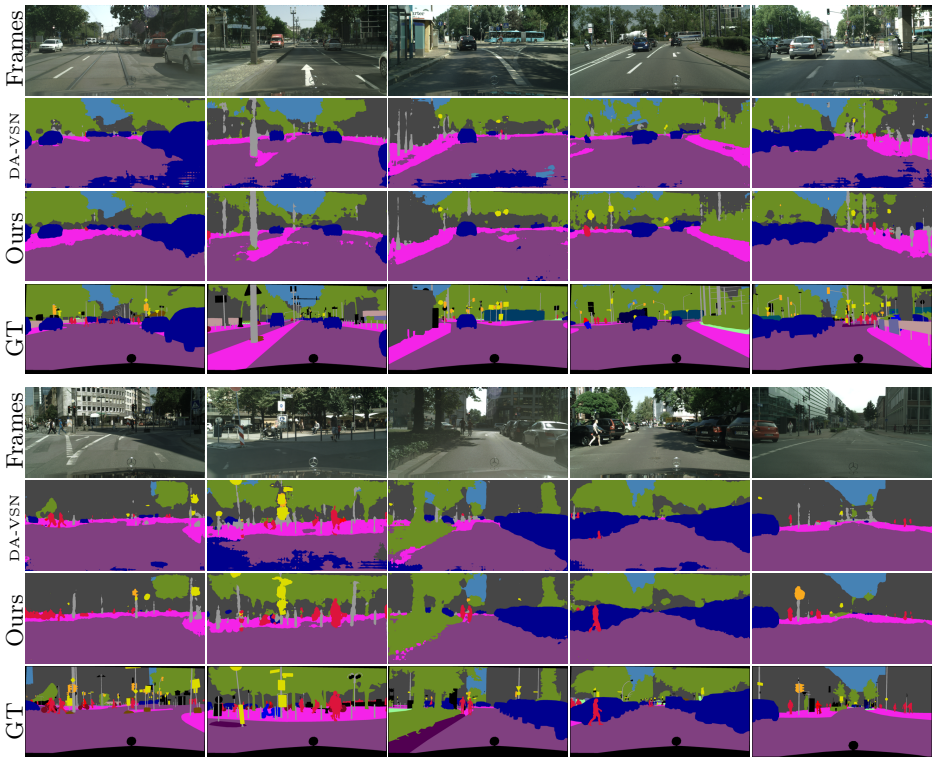

Fig. S2. Qualitative comparison results on the SYNTHIA → Cityscapes domain adaptive video segmentation task using 10 independent samples from the Cityscapes validation set.

References

1. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: ECCV. pp. 611–625. Springer (2012) 2
2. Guan, D., Huang, J., Xiao, A., Lu, S.: Domain adaptive video segmentation via temporal consistency regularization. In: ICCV. pp. 8053–8064 (2021) 2
3. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: CVPR. pp. 2462–2470 (2017) 2