A. More Dataset Details

• Cityscapes-Seq [2] is a widely used real dataset that contains 2,975 and 500 video sequences for training and evaluation, respectively. Specifically, each sequence involves 30 consecutive frames with resolution of 1024×2048, while only one frame among the sequence is fully annotated.

• SYNTHIA-Seq [6] consists of 8,000 simulated video frames with the resolution of 760×1280 and pixel-level annotations automatically produced by game engine. Similar to [3], we evaluate on the 11 classes in common with the Cityscapes-Seq.

• VIPER [5] contains 133,670 synthesized video frames with the resolution of 1080×1920. The full annotations in VIPER are available for all frames, which are collected by a virtual moving object in diverse ambient conditions. Following the setup of [3], we use the 15 classes in line with Cityscapes-Seq.

B. More Implementation Details

We provide more details here for the image augmentations we use in our experiments. The combination of augmentations for each training sample is selected randomly from the augmentation set, including color jitter (i.e. brightness, contrast, saturation and hue), gaussian blur, random flipping and scaling. For completeness, we listed the detail of the transformations in Table 1.

Table 1. List of Data Transformations

Transformation	Description	Range
Brightness	Adjust the brightness of the image	[0.2, 1.8]
Contrast	Control the contrast of the image	[0.2, 1.8]
Saturation	Adjust the saturation of the image	[0.2, 1.8]
Hue	Adjust hue of image by shifting RGB channels	[0.8, 1.2]
Gaussian Blur	Adapt Gaussian Blur to the image	{5, 7, 9}
Horizontal Flip	Flip image and label horizontally	-
Rescale	Rescale the size of image	[0.8, 1.2]

C. More Qualitative Comparisons

We qualitatively compare the proposed TPS with two best-performing baselines *DA-VSN* [3] and *Pixmatch* [4] over two domain adaptive video segmentation benchmarks. Figs. 1 and 2 show the comparisons, where three consecutive video frames are shown in each figure. It can be observed that the proposed TPS outperforms both DA-VSN and PixMatch clearly and consistently.

For further evaluation, we compare our method with the state-of-the-arts on real-scene long video sequence from Cityscapes. Instead of directly using test data that only contains short sequences (30 consecutive frames), we evaluate

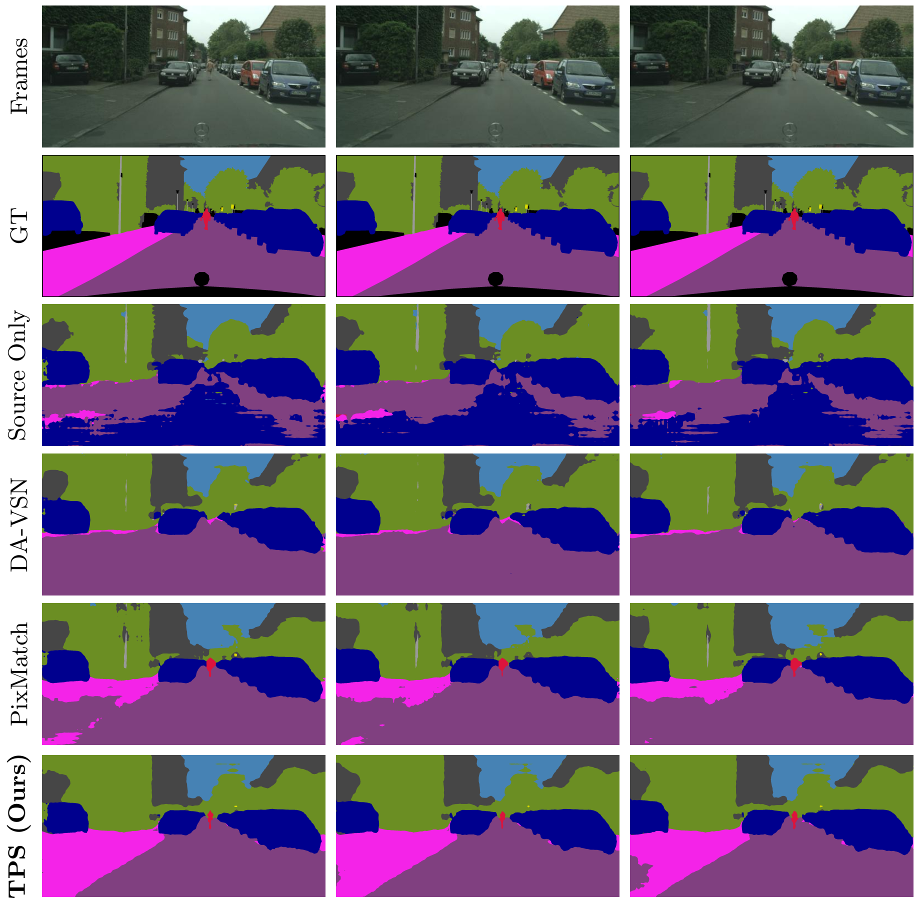

Fig. 1. Qualitative comparison of TPS with the state-of-the-art over domain adaptive video segmentation benchmark "SYNTHIA-Seq → Cityscapes-Seq": TPS produces much more accurate segmentation as compared to "source only", indicating the effectiveness of our approach on addressing domain adaptation issue. Moreover, TPS generates better segmentation than DA-VSN [3] and PixMatch [4] as shown in rows 4-5, which is consistent with our quantitative result. Best viewed in color.

our method on the Cityscapes video demo that lasts much longer (hundreds of frames each sequence, 3 sequences in total).[1] We pick one sequence for each benchmark and make further comparisons on both benchmarks (i.e. SYNTHIA-Seq→Cityscapes-Seq and VIPER→Cityscapes-Seq). The complete record is provided in https://github.com/xing0047/TPS/releases/tag/demo.

[1] https://www.cityscapes-dataset.com/file-handling/?packageID=12/

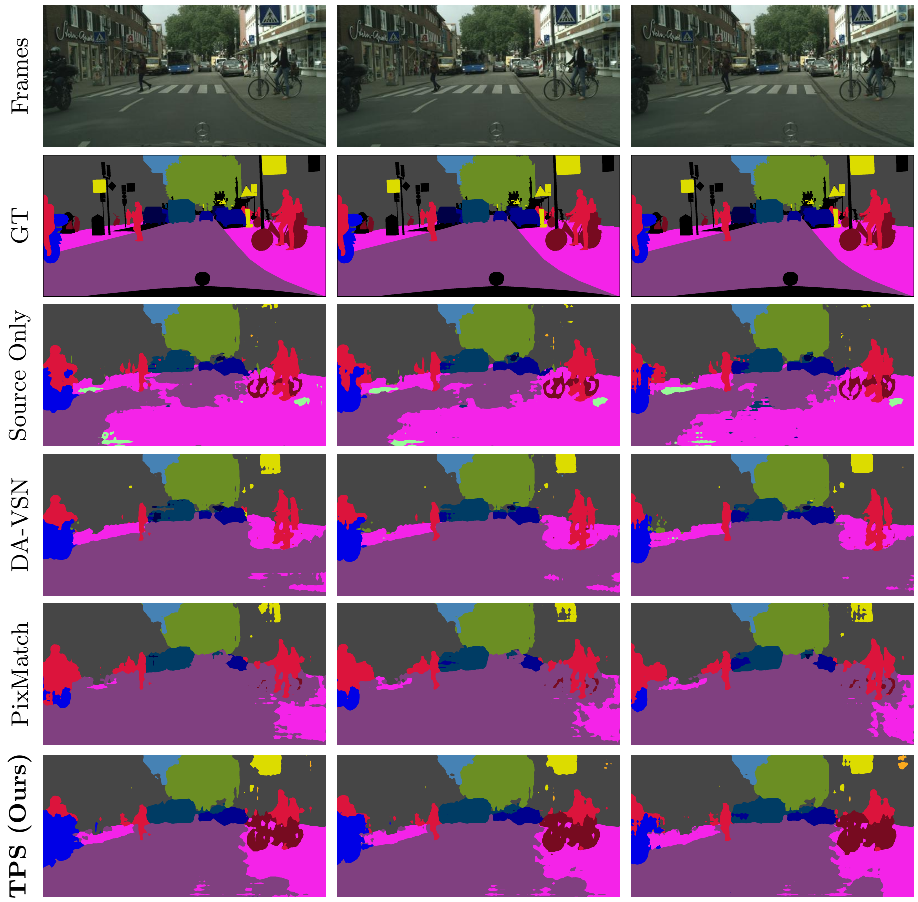

Fig. 2. Qualitative comparison of TPS with the state-of-the-art over domain adaptive video segmentation benchmark "VIPER → Cityscapes-Seq": TPS produces much more accurate segmentation as compared to "source only", indicating the effectiveness of our approach on addressing domain adaptation issue. Moreover, TPS generates better segmentation than DA-VSN [3] and PixMatch [4] as shown in rows 4-5, which is consistent with our quantitative result. Best viewed in color.

D. More Quantitative Comparisons with Consistency-training-based Methods

In the Section 4.2, we compared the proposed TPS with the state-of-the-art method on domain adaptive image segmentation using consistency training (the same learning scheme as in this work). We further reproduce recent consistency-training-based approaches SAC [1] and DACS [7] for domain adaptive image segmentation task and evaluate on both video adaptive semantic segmentation benchmarks. We note that TPS outperforms all the consistency-training-based methods in Tabs. 8 and 9, which demonstrates the superiority of our approach.

Table 2. Quantitative comparisons over the benchmark of SYNTHIA-Seq $\rightarrow$ Cityscapes-Seq: TPS outperforms multiple consistency-training-based domain adaptation methods [4, 1, 7] by large margins. Note that "Source only" denotes the network trained with source-domain data solely. Abbreviations for 'sidewalk', 'building', 'vegetation' and 'person' are noted as 'side.', 'buil.', 'vege.' and 'pers.' for simplicity

SYNTHIA-Seq → Cityscapes-Seq												
Methods	road	side.	buil.	pole	light	sign	vege.	sky	pers.	rider	car	mIoU
Source only	56.3	26.6	**75.6**	25.5	5.7	15.6	71.0	58.5	41.7	17.1	27.9	38.3
SAC [1]	87.0	41.1	64.0	20.4	12.1	32.8	38.2	47.6	53.1	19.3	81.1	48.9
DACS [7]	86.4	40.0	74.0	**27.8**	9.5	28.2	**71.6**	**72.0**	55.6	20.0	76.4	51.0
PixMatch [4]	90.2	49.9	75.1	23.1	17.4	34.2	67.1	49.9	55.8	14.0	84.3	51.0
TPS (Ours)	**91.2**	**53.7**	74.9	24.6	**17.9**	**39.3**	68.1	59.7	**57.2**	**20.3**	**84.5**	**53.8**

Table 3. Quantitative comparisons over the benchmark of VIPER $\rightarrow$ Cityscapes-Seq: TPS outperforms multiple consistency-training-based domain adaptation methods [4, 1, 7] by large margins. Abbreviations for 'sidewalk', 'building', 'vegetation', 'terrain', 'person' and 'motor' are noted as 'side.', 'buil.', 'vege.', 'terr.', 'pers.' and 'mot.' correspondingly

VIPER → Cityscapes-Seq																
Methods	road	side.	buil.	fence	light	sign	vege.	terr.	sky	pers.	car	truck	bus	mot.	bike	mIoU
Source only	56.7	18.7	78.7	6.0	22.0	15.6	81.6	18.3	80.4	59.9	66.3	4.5	16.8	20.4	10.3	37.1
DACS [7]	69.6	24.1	76.9	9.1	16.1	15.3	74.1	20.3	76.5	59.4	74.8	38.6	43.1	7.7	1.9	40.5
SAC [1]	52.2	19.6	73.4	3.7	23.1	25.2	73.9	17.3	78.1	56.9	**80.3**	38.3	**48.2**	17.8	14.1	41.5
PixMatch [4]	79.4	26.1	**84.6**	**16.6**	**28.7**	23.0	**85.0**	**30.1**	**83.7**	58.6	75.8	34.2	45.7	16.6	12.4	46.7
TPS (Ours)	**82.4**	**36.9**	79.5	9.0	26.3	**29.4**	78.5	28.2	81.8	**61.2**	80.2	**39.8**	40.3	**28.5**	**31.7**	**48.9**

References

1. Araslanov, N., Roth, S.: Self-supervised augmentation consistency for adapting semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15384–15394 (2021)
2. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
3. Guan, D., Huang, J., Xiao, A., Lu, S.: Domain adaptive video segmentation via temporal consistency regularization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8053–8064 (2021)
4. Melas-Kyriazi, L., Manrai, A.K.: Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12435–12445 (2021)
5. Richter, S.R., Hayder, Z., Koltun, V.: Playing for benchmarks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2213–2222 (2017)
6. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3234–3243 (2016)
7. Tranheden, W., Olsson, V., Pinto, J., Svensson, L.: Dacs: Domain adaptation via cross-domain mixed sampling. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1379–1389 (2021)