Domain Adaptive Video Semantic Segmentation via Cross-Domain Moving Object Mixing (Supplementary Material)

Kyusik Cho Suhyeon Lee Hongje Seong Euntai Kim*
School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
{ks.cho,hyeon93,hjseong,etkim}@yonsei.ac.kr

1. Visualization of FATC

We propose Feature Alignment with Temporal Context (FATC) to filter out unreliable predictions with temporal consensus. To show the robustness of the filtering idea, we visualize an example of the process in Figure 1. For (a) the prediction, we zoom a certain region and show (b) the current frame prediction ($p_t^{\mathbb{T}}$), (c) previous frame prediction ($p_{t-\tau}^{\mathbb{T}}$), (d) result after filtering with temporal context ($\mathbb{1}(p_t^{\mathbb{T}}, p_{t-\tau}^{\mathbb{T}})$), and (e) ground truth ($y_t^{\mathbb{T}}$). (f) Corresponding ground truth for the full image is shown for visibility. In the second row, we additionally show (g-i) the error maps. Here, blue-colored pixels indicate the correct prediction, the reds indicate the wrong predicted pixels, and the blacks indicate the removed pixels. As shown in the figure, the filtering with temporal context removes the error regions in the current frame ($p_t^{\mathbb{T}}$) effectively. Since our feature alignment algorithm generated the feature centroid with blue and red colored pixels, our filtering algorithm will contribute to make more robust feature centroid.

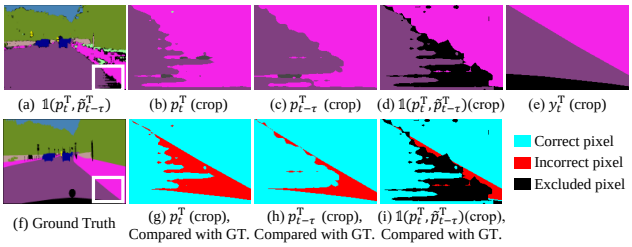

Figure 1. **Effect of FATC.** The first row shows semantic segmentation maps and the second row shows error maps. In the error map, correct pixels are colored in blue, wrong predicted pixels are colored in red, and pixels excluded by temporal consensus are colored in black.

2. Reproducibility

To show the stability and reproducibility of our methods, we present additional results of our method by training the network three times with different random seeds. We obtain the results by mIoU of 53.81, 53.53, and 54.78 (average 54.04 ± 0.54). The three experimental results show that our proposal is stable and reproducible.

3. More Quantitative Results

We further present additional qualitative results for real-world videos from Cityscapes demoVideo [1] dataset. The results are available online: `https://youtu.be/xrfe21mNQh0`

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

*Corresponding author