Supplementary Material: Temporally-Consistent Video Semantic Segmentation with Bidirectional Occlusion-guided Feature Propagation

Razieh Kaviani Baghbaderani[1] Yuanxin Li[1] Shuangquan Wang[1] Hairong Qi[2]
[1]SOC R&D, Samsung Semiconductor, Inc.
[2]The University of Tennessee, Knoxville, TN, USA
{r.kaviani, yuanxin.li, shuangquan.w}@samsung.com hqi@utk.edu

A Class-wise IoU

In order to demonstrate how much accuracy drops for each class after our feature propagation and rectification steps, the class-wise IoUs are presented in Tab. 5. One can see that most of objects are successfully propagated to non-keyframes with a tiny drop which could be negligible compared to the temporal consistency gain reported in Tab. 1 in the main paper. It should be mentioned that it is common in keyframe-based approaches that slender objects, like pole or traffic sign, get deteriorated due to the flow-based propagation.

Table 5. Class-wise IoUs on the Cityscapes validation set with HRNetV2 and DeeplabV3+ as baselines.

Class	HRNetV2	Ours	DeeplabV3+	Ours
Road	98.2	98.1	98.1	98.1
Sidewalk	85.3	84.7	84.6	84.5
Building	92.4	91.9	92.0	91.7
Wall	56.0	58.0	59.0	58.6
Fence	60.2	60.9	60.8	59.9
Pole	64.6	57.4	58.8	55.0
Traffic light	69.9	67.9	64.9	63.8
Traffic sign	77.2	76.3	75.0	73.9
Vegetation	92.5	91.9	92.1	92.0
Terrain	64.2	64.4	64.7	65.2
Sky	94.3	94.0	94.5	94.5
Person	81.2	78.4	79.6	78.8
Rider	59.4	58.5	59.0	58.7
Car	94.6	94.3	94.6	94.6
Truck	67.6	75.7	83.8	84.8
Bus	81.4	81.1	85.5	88.5
Train	67.6	69.2	70.5	73.1
Motorcycle	59.7	62.0	63.7	64.9
Bicycle	75.9	74.1	74.6	74.0
Avg.	75.9	75.7	76.6	76.5

B Visual Comparison

To compare our proposed BOFP with the keyframe and non keyframe-based methods over time, two videos showing the results simultaneously are provided.

B.1 Comparison with Keyframe-based Methods

The video consists of the original frame, video semantic segmentation results obtained from DFF [3] and DAVSS [4] as keyframe-based methods, and our proposed method. Comparing our method with the keyframe-based methods, the regions where flow-based feature propagation causes mistakes are described.

Note to the person riding a bicycle. The traffic-signs and pedestrians are being occluded as the rider passes them. It can be seen that the distortions made by flow-based feature propagation can be corrected by our method as these occluded regions are compensated with the features propagated from the subsequent key frame.

B.2 Comparison with non Keyframe-based Methods

The video consists of the original frame, video semantic segmentation results obtained from HRNetV2 [2] and GRFP [1] as non keyframe-based methods, and our proposed method.

Comparing our method with the non keyframe-based methods, it can be observed that there are regions that HRNetV2 and GRFP methods produce predictions which are not consistent over time. Hence, the results suffer from the flickering problem leading to inferior temporal consistency.

References

[1] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6819–6828, 2018. 1

[2] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui

Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1

[3] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2349–2358, 2017. 1

[4] Jiafan Zhuang, Zilei Wang, and Bingke Wang. Video semantic segmentation with distortion-aware feature correction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. 1