Event-guided Low-light Video Semantic Segmentation: Supplementary Materials

Zhen Yao
Lehigh University
zhy321@lehigh.edu

Mooi Choo Chuah
Lehigh University
chuah@cse.lehigh.edu

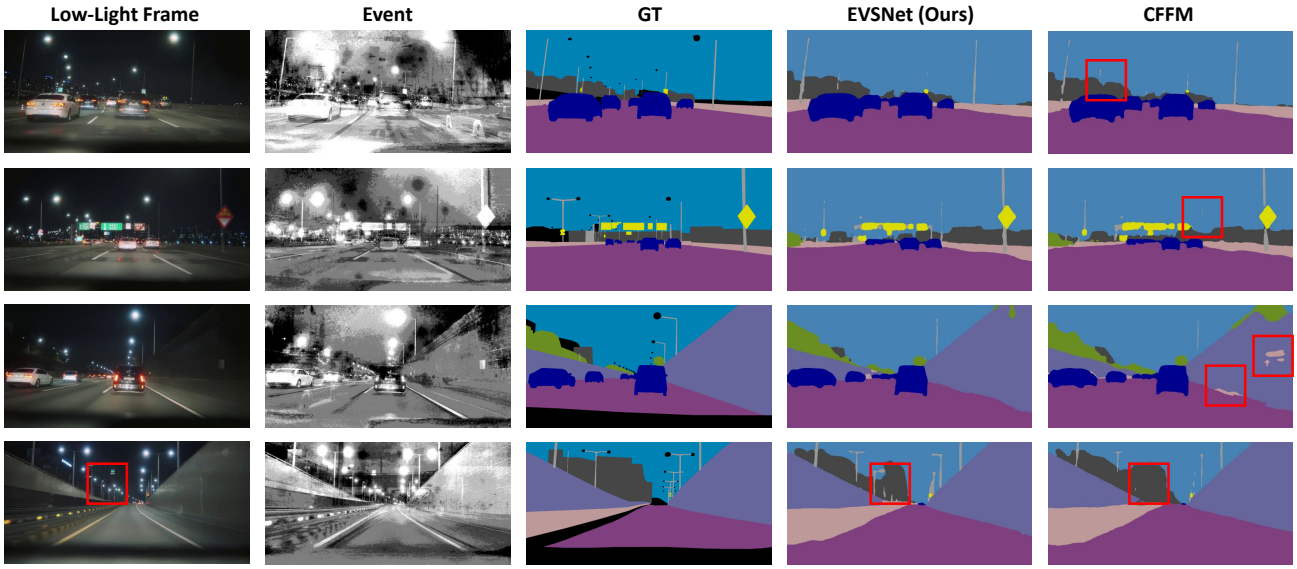

Figure 1. Qualitative results on NightCity dataset [5]. It shows that our model generates more robust and temporal consistent results, compared to the SOTA method. Best viewed in color.

1. Training Details

Loss function. Our main loss is standard pixel-wise Cross-Entropy loss. Following common practice, two auxiliary losses are also computed: (1) we attach an auxiliary Fully Convolutional Networks (FCN) [2] head upon the third-stage features to generate segmentation results at 1/8 resolution and apply Cross-Entropy loss to calculate an auxiliary loss; (2) we use an auxiliary MLP head upon the multi-scale features after the Multi-scale Mixer to optimize per-frame learning. We train our model by optimizing the main loss and two auxiliary losses jointly using a weighted sum, which is set to 1.0, 0.4, and 0.4, respectively:

2. Additional Ablation Study

Selection of Temporal Decoder. Besides, we also show the results of different selections of the Temporal Decoder module in Table 1 using MiT-B0 backbone on the low-light VSPW [3] dataset. We start by concatenating features from all frames and using a simple MLP decoder to predict. When replacing it with a more sophisticated decoder, such as a Temporal block in Fiery [1] or a Focal block in Focal Transformer [6], results are better.

Table 1. Ablation study of multiple selections of Temporal Decoder

Temporal Decoder	mIOU↑	mVC_8 ↑	mVC_{16} ↑
Concatenation	23.3	83.8	77.7
Temporal block [1]	25.4	84.9	79.2
Focal block (ours) [6]	28.2	87.0	82.1

3. Additional Qualitative Results

We visualize segmentation predictions on several video frames from the NightCity [5] dataset to demonstrate the robustness of our proposed model in real-world scenarios, as

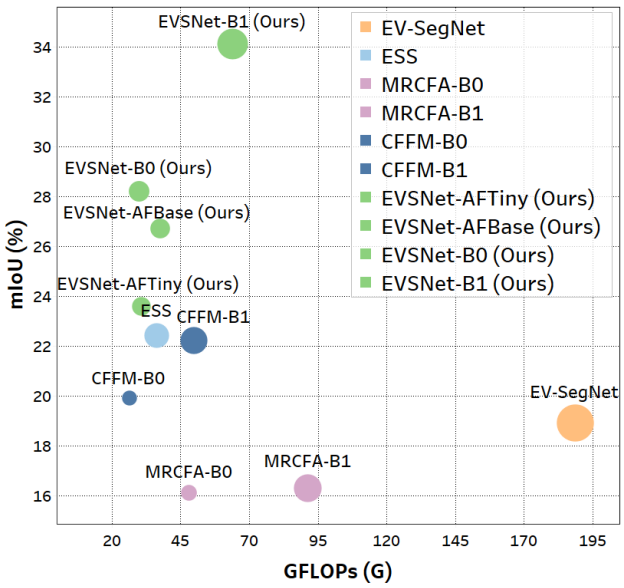

Figure 2. Performance vs. computational cost on the low-light VSPW dataset. The circle areas are proportional to the model parameter size.

shown in Figure. 1. We compare the qualitative results of our method with the baseline CFFM [4] using its default settings. In CFFM's predictions, some small objects are omitted, showing its struggle to recognize motions of small regions. EVSNet generates more accurate boundaries and captures the temporal motions, demonstrating the effectiveness of EVSNet.

Compared with the difference in predictions of low-light VSPW [3] between EVSNet and CFFM, EVSNet doesn't improve significantly. That is because (1) authors manually sample frames from the videos in the NightCity dataset, leading to video frames being inconsistent and further affecting the quality of event frames; (2) label annotations are incomplete, e.g. fourth image in Figure. 1. The authors failed to annotate all the poles in that image, but our model detects and segments them.

4. Performance vs. computational cost

We additionally report the Performance vs. computational cost analysis. Fig. 2 shows the EVSNet improves the SOTA performance by a large margin while slightly increasing the model size. In specific, our model only increases the GFLOPs by 14% (MiT-B0) and 28% (MiT-B1) than the baseline (CFFM) and the parameter size is about 4 Mb more than the baseline. Compared to slightly larger model size, our model increases the mIoU by 42% (MiT-B0) and 54% (MiT-B1).

References

[1] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15273–15282, 2021.

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[3] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4133–4143, 2021.

[4] Guolei Sun, Yun Liu, Henghui Ding, Thomas Probst, and Luc Van Gool. Coarse-to-fine feature mining for video semantic segmentation. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3126–3137, 2022.

[5] Xin Tan, Ke Xu, Ying Cao, Yiheng Zhang, Lizhuang Ma, and Rynson WH Lau. Night-time scene parsing with a large real dataset. *IEEE Transactions on Image Processing*, 30:9085–9098, 2021.

[6] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021.