Supplementary Material—CMDA: Cross-Modality Domain Adaptation for Nighttime Semantic Segmentation

Ruihao Xia[1] Chaoqiang Zhao[1] Meng Zheng[2] Ziyan Wu[2] Qiyu Sun[1] Yang Tang[1*]
[1]East China University of Science and Technology [2]United Imaging Intelligence
{xia_rho, zhaocq, qysun}@mail.ecust.edu.cn, {meng.zheng, ziyan.wu}@uii-ai.com
yangtang@ecust.edu.cn

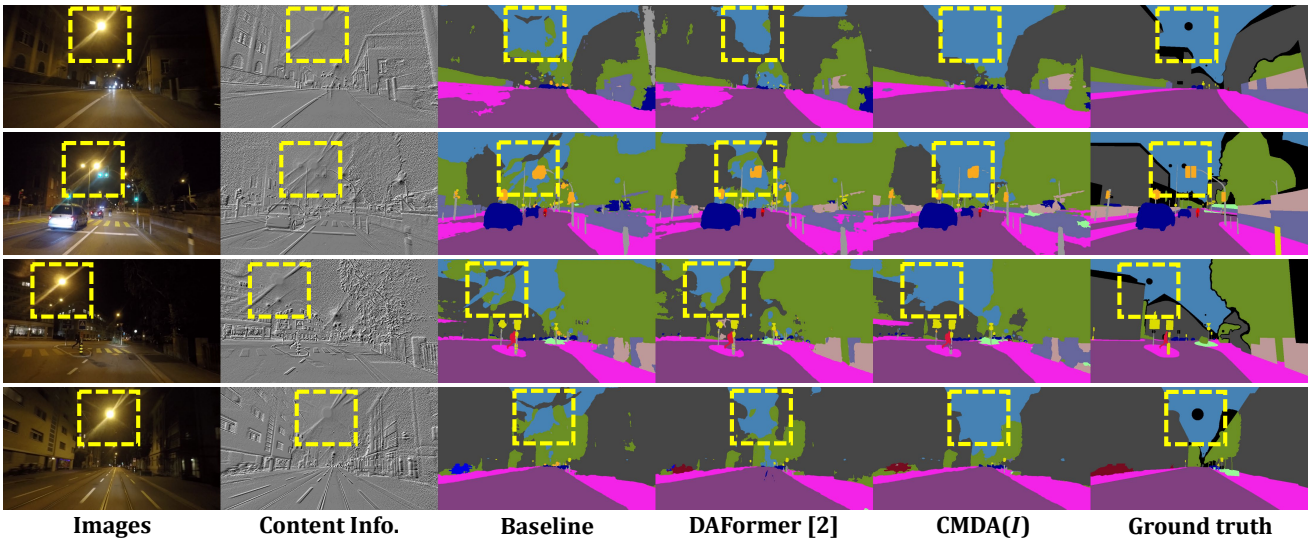

Figure 1. Qualitative results of our baseline, SOTA approach DAFormer [2], and our proposed CMDA(I) in the image-based Dark Zurich dataset [6]. Note that the content information generated by our proposed Image Content-Extractor are only utilized during training in CMDA(I).

A. Event Representation

The event camera outputs a continuous stream of events, wherein each event consists of four distinct elements, namely (t, x, y, p). Here, t denotes the trigger time, (x, y) represents the spatial coordinate, and $p \in \{+1, -1\}$ is the polarity that represents the sign of the brightness change [3].

Raw events are discrete spatial-temporal points that pose challenges for feature extraction and integration with image modalities. To overcome this, we follow the previous approach [8] to embed raw events as an image $E \in \mathbb{R}^{H \times W \times B}$, where B represents the number of temporal bins. A higher value of B indicates a more refined representation of temporal information. However, in our proposed CMDA, we focus on the High Dynamic Range (HDR) of the event camera instead of the high temporal resolution. Moreover, to ensure consistency in the number of channels of E_{ME} and E for training the style transfer network $G_{E_{ME} \to E}$, we set $B = 1$.

B. Annotations Distribution

Our proposed DSEC Night-Semantic dataset contains 18 classes. Distribution of annotations across individual classes is provided in Figure 3.

C. Training details

Style Transfer Network $G_{E_{ME} \to E}$. Following CycleGAN [9], we randomly select 1,000 E_{ME} and E_t from the Cityscapes and DSEC dataset. Then, cycle consistency and adversarial loss are utilized to train the network for 200 epochs.

Data Augmentation. In the source domain, namely the Cityscapes [1] dataset, we resize I_s, $\hat{E}_s$, and I_{CE_s} to

*Corresponding author.

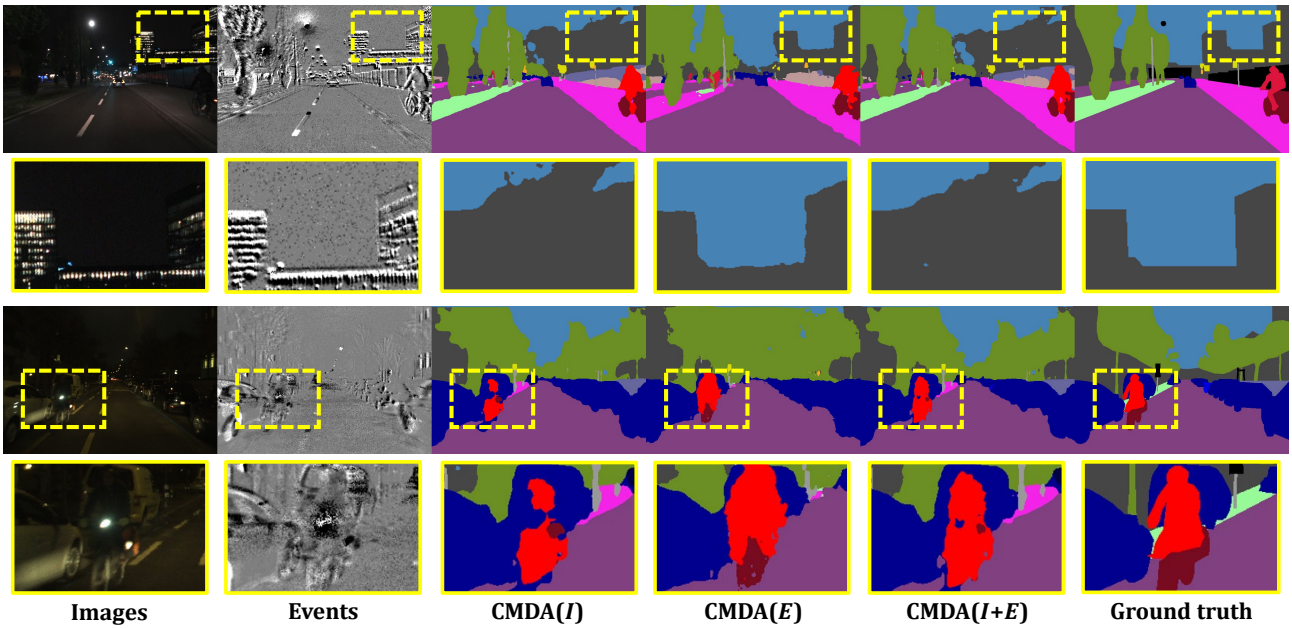

Figure 2. The failure cases of our CMDA in the proposed DSEC Night-Semantic image-event dataset.

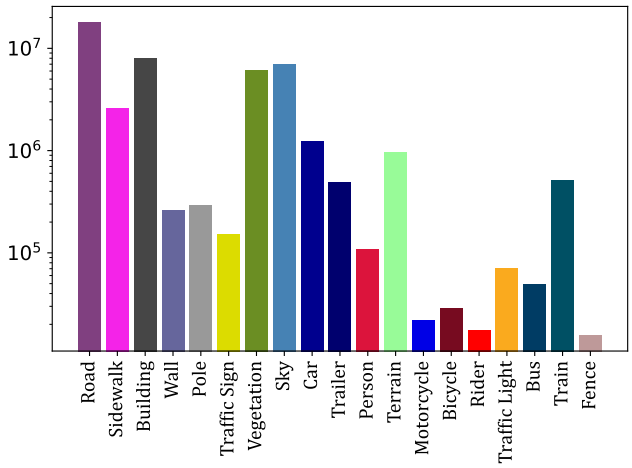

Figure 3. Number of annotated pixels (y-axis) per classes (x-axis) for our proposed DSEC Night-Semantic dataset.

1024×512 and randomly crop them into 512×512, as per the DAFormer [2].

For the target domain, namely our proposed DSEC Night-Semantic dataset, we randomly crop areas of 400×400 on I_t, E_t, and I_{CE_t} and resize them to 512×512.

During the calculation of the target loss $\mathcal{L}_t$, we follow DACS [7] and apply additional data augmentation techniques, *i.e.,* color jitter, Gaussian blur, and ClassMix [4], on the input images I_t of f^S. The corresponding I_{CE_t} are directly generated from I_t with the proposed Image-Content Extractor, while E_t are exclusively enhanced by ClassMix [4].

D. Visualization in Dark Zurich

In this section, we demonstrate the performance of the proposed Image Content-Extractor in the image-based Dark Zurich dataset [6], and compare it with the SOTA approach DAFormer [2]. As shown in Figure 1, our proposed Image Content-Extractor effectively mitigates the impact of nighttime glare, resulting in clearer edge segmentation of the sky and other objects.

E. Failure Cases

Our proposed CMDA integrates the event modality into nighttime semantic segmentation for the first time, leading to a significant improvement in segmentation performance. However, our CMDA may fail to generate satisfactory results in some cases. We compare these results with different modalities inputs in Figure 2.

Looking at the event modality in the first row, it is evident that the HDR of nighttime events provides a clear contrast between the edges of buildings. Consequently, the building and the sky in the yellow box of CMDA(E) are accurately segmented with event inputs. However, when fusing images with events, CMDA($I + E$) failed to fully utilize the benefits of the event modality. The results in the second row shows a similar situation, where events capture more robust features in the corner cases, yet CMDA fails to integrate events effectively.

The aforementioned cases show that our CMDA puts higher weights on image modality during fusion, which results in the under-utilization of event modality. We attribute this to the fact that in the source domain, daytime images

typically contain a vast majority of favorable information in the scene. As a result, CMDA can generate satisfactory segmentation results even when just relying on the image modality, and the weights of the event modality is lowered. Conversely, in nighttime scenes, event modality demonstrates its HDR advantage. Nonetheless, due to the absence of ground truth for supervised training, pseudo labels generated by f^T tend to rely more on the image modality.

F. Limitations

- **Paired Images and Events.** Our CMDA requires nighttime paired event and image modalities for training, so we wrap the 1440×1080 images to the 640×480 event coordinates. However, this operation compromises the advantage of high-resolution in the original images, and may have a negative impact on fine-grained segmentation. Therefore, future studies could focus on how to directly fuse unpaired image and event modalities, thereby leveraging the high resolution of images and HDR nighttime events.

- **Short-Time Events.** Our CMDA employs events captured within 50ms window as input. However, it is worth noting that short-time events may not provide a comprehensive representation of the scene, particularly when the relative motion between the scene and the event camera is weak. This is precisely why we choose to fuse events with images rather than relying solely on events. Therefore, future studies could explore approaches to express a comprehensive representation of the scene by utilizing events over an extended time range.

- **Reliability of Generated Events $\hat{E}_s$.** The ESIM events simulator [5] guarantees the high temporal resolution of events by interpolating a large number of frames between two adjacent images. Conversely, our proposed Image Motion-Extractor utilizes only the difference between two images to simulate events, which undoubtedly ignores the temporal information of the events and generates unreliable events compared to real events. However, in this paper, we focus mainly on the high dynamic range advantages provided by the event modality. Thus, events are embedded as a single channel image form and the temporal information is discarded. Future studies could explore the impact of high temporal resolution of events on nighttime semantic segmentation.

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. 1

[2] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022. 1, 2

[3] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008. 1

[4] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. ClassMix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1378, 2021. 2

[5] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: An open event camera simulator. In *Conference on Robot Learning*, pages 969–982. PMLR, 2018. 3

[6] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7374–7383, 2019. 1, 2

[7] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. DACS: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021. 2

[8] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 1

[9] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 1