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

Abstract

Most nighttime semantic segmentation studies are based on domain adaptation approaches and image input. However, limited by the low dynamic range of conventional cameras, images fail to capture structural details and boundary information in low-light conditions. Event cameras, as a new form of vision sensors, are complementary to conventional cameras with their high dynamic range. To this end, we propose a novel unsupervised Cross-Modality Domain Adaptation (CMDA) framework to leverage multi-modality (Images and Events) information for nighttime semantic segmentation, with only labels on daytime images. In CMDA, we design the Image Motion-Extractor to extract motion information and the Image Content-Extractor to extract content information from images, in order to bridge the gap between different modalities (Images $\rightleftharpoons$ Events) and domains (Day $\rightleftharpoons$ Night). Besides, we introduce the first image-event nighttime semantic segmentation dataset. Extensive experiments on both the public image dataset and the proposed image-event dataset demonstrate the effectiveness of our proposed approach. We open-source our code, models, and dataset at https://github.com/XiaRho/CMDA.

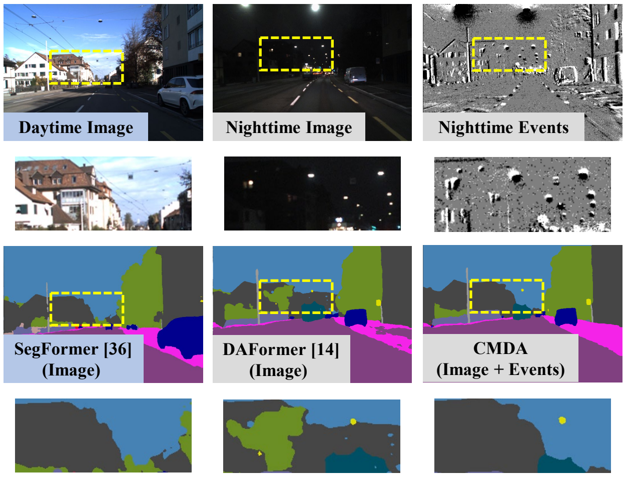

Figure 1. Images captured at different moments in the same location show that the low dynamic range of frame-based cameras leads to reduced color contrast and detailed edges of objects at night. To overcome this challenge, we introduce event cameras that have a high dynamic range and are capable of capturing more nighttime details. In comparison to the semantic segmentation results obtained from daytime images [36], nighttime images result in misclassification cases [14]. However, our proposed CMDA improves this by introducing event modality for the first time.

1. Introduction

Semantic segmentation is a crucial aspect of computer vision, which is essential for many applications, such as autonomous driving [20, 28], robotics [4, 19, 21], and surveillance [18]. While semantic segmentation of daytime scenes has made significant progress [5, 29, 36, 41], challenges remain unsolved for nighttime scenes due to the much-degraded image quality at night, as well as the lack of high-quality annotations. Most existing works [11, 33, 34, 37] employed unsupervised domain adaptation (UDA) for nighttime semantic segmentation to solve the label scarcity problem, which leverage labeled daytime images (Source Domain) and unlabeled nighttime images (Target Domain). However, the low dynamic range of conventional frame-based cameras results in poor image quality at night compared to daytime images, *i.e.,* the decrease in color contrast and details results in a reduction of clarity in nighttime images. This impedes the effective discrimination of object boundaries. Thus, the performance of methods solely relying on nighttime images as input is limited.

To address the limitations of frame-based cameras, we propose to employ event cameras for nighttime semantic segmentation. Event cameras output the spatio-temporal coordinates of pixels whose luminosity changes exceeding a certain threshold value [9, 17]. Their unique operating principle offers a higher dynamic range (140 dB vs. 60 dB)

[*] Corresponding author.

over frame-based cameras [10], which enhances contrast in low-light scenarios, facilitating more precise segmentation of objects. On the other hand, events are asynchronous and spatially sparse, lacking a comprehensive representation of the scene. Hence methods based solely on events are typically inferior to image-based approaches [31, 32]. To this end, we propose the first image-event cross-modality framework, Cross-Modality Domain Adaptation (CMDA), to leverage both image and event modalities for nighttime semantic segmentation in an unsupervised manner. As shown in Figure 1, compared to conventional image-based UDA approaches, our framework achieves substantially improved nighttime semantic segmentation performance with the combination of event modality.

In the proposed CMDA, the key challenges lie in establishing the connection between image and event modalities, as well as minimizing the domain shifts between the representations of daytime and nighttime images. Specifically:

Challenge 1: Images $\rightleftharpoons$ Events. The absence of event modality in the source domain hinders the fusion of images and events. An intuitive idea is to transfer the daytime images into events. However, event cameras record the movement of the scene w.r.t. the camera, which cannot be determined with a single image. Thus, we propose the Image Motion-Extractor to extract motion information from adjacent images and bridge the gap between image and event modalities.

Challenge 2: Day $\rightleftharpoons$ Night. Images can typically be separated into content and style information [16]. Previous image-based UDA approaches employed a style transfer network [44] to transform daytime images so they look like nighttime [11, 37]. However, the transferred images are often unrealistic and unreliable, due to the significant and heterogeneous noise at night [40]. In contrast, we eliminate daytime and nighttime style information and preserve only content information based on the proposed Image Content-Extractor, which transfers both daytime and nighttime images to a common content domain.

Then, we construct our network based on the image-based UDA method DAFormer [14]. Instead of taking only images as input, we combine events with images to perform improved nighttime semantic segmentation, with domain adaptation from labeled daytime images. In addition, as there are no existing benchmark datasets in the community for nighttime image-event semantic segmentation evaluation, we follow the image-based Dark Zurich dataset [25] and manually annotate 150 image-event with fine, pixel-level labels from DSEC dataset [13].

In summary, our contributions are as follows:

- 1) To the best of our knowledge, we introduce the first method to utilize event modality in nighttime semantic segmentation.
- 2) We propose a novel CMDA framework by fusing image and event modalities in an unsupervised manner with only labeled images from the source domain.
- 3) We propose the Image Motion-Extractor and Image Content-Extractor to bridge the gaps between modalities (Images $\rightleftharpoons$ Events) and domains (Day $\rightleftharpoons$ Night).
- 4) To fill in the missing evaluation criteria for nighttime image-event semantic segmentation, we align images and event modalities in the DSEC dataset [13] and manually annotate 150 image-event with fine, pixel-level labels.
- 5) We show the effectiveness of our CMDA framework, which achieves SOTA results on both the existing nighttime images benchmark dataset [25] and our proposed image-event dataset.

2. Related Work

2.1. Event-based Semantic Segmentation

The problem of event-based semantic segmentation is under-explored, compared to image-based semantic segmentation due to the absence of high-quality datasets. Considering the paired image-event data in the DDD17 dataset [2], Alonso *et al.* [1] utilize a pretrained image-based network to generate pseudo labels for corresponding events. Then, labeled events data are employed to train an event-based network in a supervised manner.

Considering the supervision on intermediate features, Wang *et al.* [32] utilize a pretrained image-based teacher network for cross-modality knowledge distillation. Additionally, the training of the event-based network is aided by source data from another dataset [6]. Furthermore, Wang *et al.* [31] incorporate the cross-task knowledge transfer through an image reconstruction network to transfer the feature-level and prediction-level information. Unlike previous studies, Sun *et al.* [30] employ a pretrained recurrent network, originally designed for image reconstruction [23], to encode events and generate semantic segmentation results. However, the recurrent network requires a large number of events during both training and testing.

Datasets. Most of the existing event-based semantic segmentation datasets are synthetic datasets, *e.g.*, EventScape [12], DELIVER [38], and DADA-seg [39]. They are generated using simulators [8] or pretrained networks [42], resulting in large domain shifts compared with real-world events.

Other datasets like DDD17 [2] and DSEC [13] record real-world events, but their semantic labels are generated by pretrained image-based networks [1, 30] and only contain daytime scenes. Conversely for the first time, labels in nighttime scenes in our proposed DSEC Night-Semantic dataset are annotated manually.

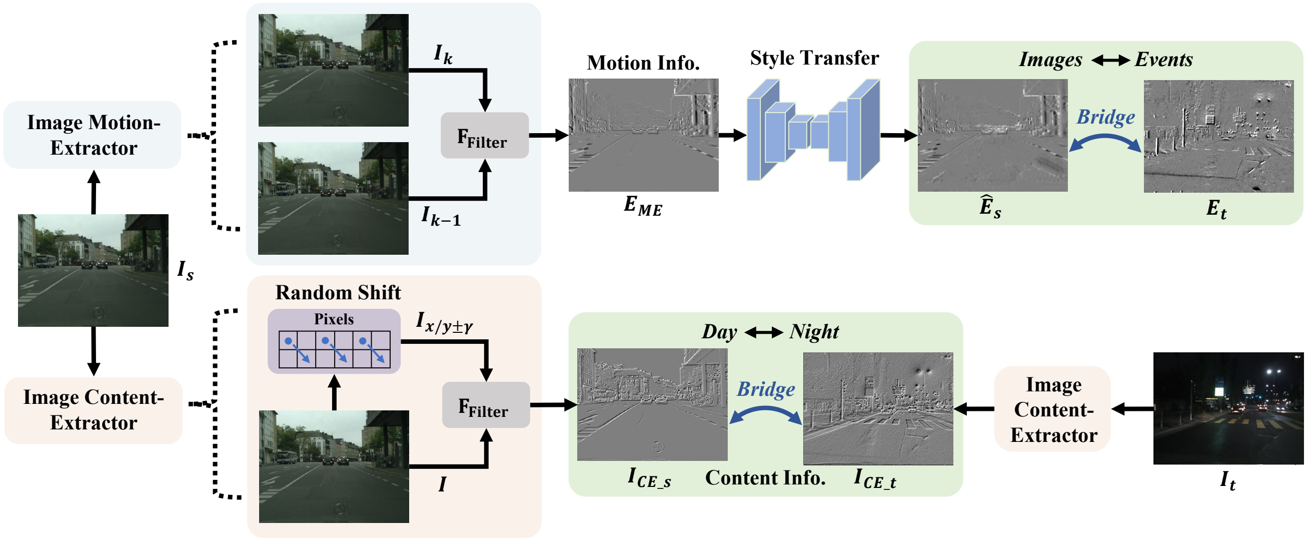

Figure 2. Processed by Image Motion-Extractor and Image Content-Extractor, E_{ME} and $I_{CE\_s/t}$ are utilized to bridge the gaps of different modalities (Images $I \rightleftharpoons$ Events E) and domains (Source Daytime $s \rightleftharpoons$ Target Nighttime t).

2.2. Nighttime Semantic Segmentation

Earlier approaches transfer daytime semantic knowledge to nighttime images via twilight images from different time periods [7] or day-to-night style transfer networks [24]. Then, the introduction of the paired day-night images dataset Dark Zurich [25] propels advancements in this task. Sakaridis *et al.* [26] transfer the labeled daytime dataset to twilight and night, utilizing curriculum learning to adapt to the unlabeled night domain. Moving away from intermediate domains and models, Wu *et al.* [33, 34] introduce an image relighting network and apply adversarial training. Xu *et al.* [37] combine the inter-domain style adaptation and intra-domain gradual self-training to achieve smooth semantic knowledge transfer. From the perspective of illumination and datasets differences, Gao *et al.* [11] propose a novel domain adaptation framework via cross-domain correlation distillation. However, paired day-night images are difficult to acquire in practical settings. Recently, the emergence of transformer brings a huge boost to nighttime semantic segmentation, and our approach falls into this category. These Transformer-based methods [14, 15] employ self-training and consistency training to achieve superior performance without the need for paired data, which have achieved SOTA performance.

However, day-to-night style transfer in Transformer-based methods leads to negative transfer, which is caused by the unrealistic and unreliable transferred nighttime images. Our proposed Image Content-Extractor transfers both domains to a shared content domain to alleviate the above issue. Then, we introduce event modality to make up for the low dynamic range of image modality for the first time.

3. Cross-Modality Domain Adaptation (CMDA)

In CMDA, given labeled images from the source domain $\{(I_s, Y_s)\}$ and unlabeled image-event pairs from the target domain $\{(I_t, E_t)\}$, our objective is to train a network f that can accurately predict segmentation masks for the image-event pair input in the target domain, *i.e.,* $f : (I_t, E_t) \to Y_t$. As there are no labels in the target domain, the key problem is to bridge the gaps between I_s and (I_t, E_t). Therefore, we design the Image Motion-Extractor to extract the motion information recorded by event cameras from I_s. Also, the Image Content-Extractor is designed to filter the style information and obtain the content information from both I_s and I_t. In the following sections, we first introduce the key components of CMDA, *i.e.,* the Image Motion-Extractor and Image Content-Extractor, followed by detailed explanations of CMDA structure as well as the training process.

3.1. Image Motion-Extractor

The absence of event data in the source domain impedes the network to associate images with events. Considering that events are represented by the relative motion between the camera and the scene, directly transferring images to events is non-trivial due to the lack of motion information in a single image. To overcome this challenge, we propose the Image Motion-Extractor to obtain the relative motion information E_{ME} from two temporally adjacent images, as illustrated at the top of Figure 2.

Considering the event camera that records the logarithmic intensity change of pixels [10], we simulate this by differencing the same pixel of two adjacent images on the logarithmic domain. Thus, given by two temporally adjacent grayscale images $I_{k-1}, I_k \in \mathbb{R}^{H\times W\times 1}$, we compute

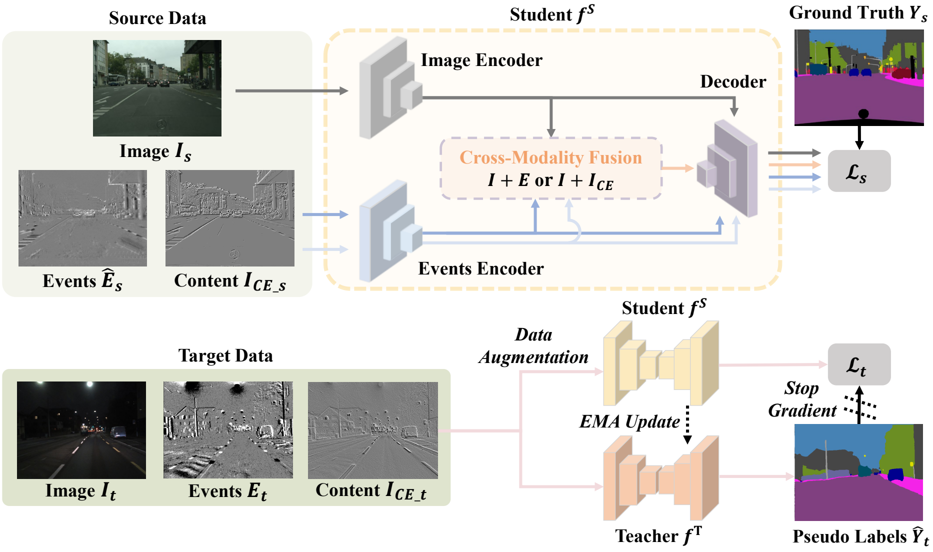

Figure 3. Two regularizations are employed to train the network: the supervised loss $\mathcal{L}_s$ in the source domain and the unsupervised domain adaptation loss $\mathcal{L}_t$ in the target domain. All losses are calculated on the student network f^S. The teacher network f^T is used to generate pseudo labels for target data and updated with the EMA of f^S.

$E_{ME} = \mathrm{F}_{\mathrm{Filter}}(I_{k-1}, I_k)$ with the following:

$$\mathrm{F}_{\mathrm{Filter}}(I_1, I_2) = \mathrm{F}_{\mathrm{Norm}}(\mathrm{F}_{\mathrm{ClipIgn}}(\mathrm{F}_{\mathrm{LogDiff}}(I_1, I_2))), \quad (1)$$

$$\mathrm{F}_{\mathrm{LogDiff}}(I_1, I_2) = \ln{(I_1 + \epsilon)} - \ln{(I_2 + \epsilon)}, \quad (2)$$

$$\mathrm{F}_{\mathrm{ClipIgn}}(x) = \min{(|x|, \alpha)} \cdot \mathrm{sgn}(x) \cdot \mathbb{1}(|x| > \beta), \quad (3)$$

$$\mathrm{F}_{\mathrm{Norm}}(x) = 2 \cdot \frac{x - \min(x)}{\max(x) - \min(x)} - 1, \quad (4)$$

where $\mathrm{F}_{\mathrm{LogDiff}}(I_1, I_2)$ represents the difference of I_1, I_2 in the logarithmic domain, ϵ is a small scalar constant to prevent taking the logarithm of zero. $\mathrm{F}_{\mathrm{ClipIgn}}(x)$ aims to clip larger values and ignore smaller values through two hyper-parameters α and β, $\mathbb{1}(\cdot)$ is the indicator function, and $\mathrm{sgn}(\cdot)$ is the signum function. $\mathrm{F}_{\mathrm{Norm}}(x)$ is the min-max normalization, scaling the values from -1 to 1.

However, like frame-based cameras, event cameras are also suffering from noise at night. To further narrow the gap between E_{ME} and E_t, we train a style transfer network [44] $G_{E_{ME} \to E}$ in an unsupervised manner to add the style of E_t to E_{ME}, resulting in transferred daytime events $\hat{E}_s = G_{E_{ME} \to E}(E_{ME})$. So far, we associate I_s with E_t with our proposed Image Motion-Extractor and $G_{E_{ME} \to E}$.

3.2. Image Content-Extractor

Previous image-based UDA approaches transferred daytime images I_s to the nighttime style with a style transfer network [44] to alleviate domain gaps [11, 37]. However, the real nighttime style is difficult to construct due to the complex and changing nighttime scenes [40]. Instead, we propose the Image Content-Extractor to obtain the content information. By eliminating the daytime and nighttime style, we transfer both I_s and I_t to the intermediate domain and discard the nighttime style generating and utilization of style transfer network.

Given a grayscale image I, we shift it γ pixels to the left/right and up/down randomly and obtain $I_{x\pm\gamma}$ and $I_{y\pm\gamma}$. Then, the intermediate shared content domain I_{CE} is generated by the following:

$$I_{CE} = \frac{1}{2} \cdot \mathrm{F}_{\mathrm{Filter}}(I, I_{x\pm\gamma}) + \frac{1}{2} \cdot \mathrm{F}_{\mathrm{Filter}}(I, I_{y\pm\gamma}) \quad (5)$$

By subtracting the shifted version of the image from itself, pixels of the same color are erased, leaving only the pixels at the edges of the scene, *i.e.,* content information.

We process both I_s and I_t to obtain I_{CE_s} and I_{CE_t}. As shown in Figure 2, after converting I into I_{CE}, the domain-specific texture (Style Information) is largely eliminated, and only the domain-invariant structure (Content Information) is retained.

3.3. Network Details

The proposed extractors mentioned above enable us to bridge the gaps between modalities and domains at the input level. In this section, we elaborate on how to effectively utilize I, E and I_{CE} within the CMDA framework.

Overview. Our CMDA is based on the image-based self-training method DAFormer [14]. The framework comprises a student network f^S and a teacher network f^T. Given source and target data as inputs, f^S outputs predicted semantic segmentation results P. These results are then computed with the source ground truth and target pseudo labels to obtain the cross-entropy loss. f^T aims to provide pseudo labels

Algorithm 1 Training of CMDA

```
Require: Source data $\{(I_s, Y_s)\}$, Target data $\{(I_t, E_t)\}$.
Obtain $E_{ME}$, $I_{CE\_s}$, and $I_{CE\_t}$ based on Eqn. (1) and Eqn. (5).
Train $G_{E_{ME}\to E}$ and obtain $\hat{E}_s = G_{E_{ME}\to E}(E_{ME})$.
Initialize $f^S$ and $f^T$ with the same pretrained network.
for $n = 1$ to $40k$ do
    Compute source loss $\mathcal{L}_s$ based on Eqn. (6).
    Generate pseudo labels $\hat{Y}_t$ by randomly choosing $E$ or $I_{CE}$ to fuse with $I$.
    Compute target loss $\mathcal{L}_t$ based on Eqn. (6).
    Loss back-propagation and update $f^S$.
    Update $f^T$ based on the EMA in Eqn. (8).
end for
```

in the target domain and is updated with the exponentially moving average (EMA) of f^S.

Network Architecture. As shown in Figure 3, both f^S and f^T consist of two encoders, one cross-modality fusion module, and one decoder. Given $I/E/I_{CE}$, the image encoder extracts the features from I, while the events encoder extracts the features from both E and I_{CE}. The fusion module is utilized to combine features from I and E/I_{CE}. Finally, the decoder receives both the fused and non-fused features and generates predicted semantic segmentation outputs P_I, P_E, $P_{I_{CE}}$, and $P_{I+E}/P_{I+I_{CE}}$.

Fusion Module. Both the image and events encoders in our framework generate features with four different scales. To fuse features from the same scale, we individually input them into the attention block adapted from SegFormer [36] and average them to obtain the fused features.

Random Choice of E or I_{CE}. To take full advantage of E as well as I_{CE} modalities, pseudo labels in the target domain are generated by fusing I with E or I_{CE} randomly, *i.e.,* $\hat{Y}_t = f^T(I_t, E_t/I_{CE\_t})$.

Training Loss. Given daytime modalities I_s, $\hat{E}_s$, I_{CE_s}, and nighttime modalities I_t, E_t, I_{CE_t}, we train the student network f^S with a combination of several categorical cross-entropy (CE) losses $\mathcal{L}_{s/t}$ calculated with daytime ground truth Y_s and nighttime pseudo labels $\hat{Y}_t$. For brevity, we omit the domain term s/t of P and Y in the following:

$$\begin{aligned}\mathcal{L}_{s/t} = {} & \lambda_I \mathcal{L}_{ce}(P_I, Y) + \lambda_E \mathcal{L}_{ce}(P_E, Y) \\ & + \lambda_{I_{CE}} \mathcal{L}_{ce}(P_{I_{CE}}, Y) \\ & + \lambda_{Fusion} \mathcal{L}_{ce}(P_{I+E}, Y), \end{aligned} \tag{6}$$

$$\mathcal{L}_{ce}(P, Y) = \sum_{j=1}^{H\times W} \sum_{c=1}^{C} Y^{(j,c)} \log \delta(P^{(j,c)}), \tag{7}$$

where $\delta(P)$ denoted the softmax output of the predicted results P, C is the number of semantic classes, λ_I, λ_E, $\lambda_{I_{CE}}$, and λ_{Fusion} are hyper-parameters.

Sequence	Training samples	Testing samples
Zurich City 09a	508	45
Zurich City 09b	109	9
Zurich City 09c	371	34
Zurich City 09d	478	42
Zurich City 09e	226	20
Total	**1,692**	**150**

Table 1. The dataset split of our proposed DSEC Night-Semantic dataset.

In contrast to f^S, which is updated through gradient descent, f^T is updated by the exponentially moving average (EMA) of the weights of f^S in each training step following DAFormer [14]:

$$f^T = \sigma f^T + (1-\sigma) f^S, \tag{8}$$

where σ is a momentum parameter.

We summarize the overall training process of our CMDA framework in Algorithm 1.

4. Experiments

4.1. Implementation Detail

Our baseline model is adopted from DAFormer [14] without the loss of Thing-Class Feature Distance. Building upon this baseline, we incorporate an events encoder and a cross-modality fusion module into the network structure. For loss weighting, we use $\lambda_I = \lambda_{Fusion} = 0.5$ and $\lambda_E = \lambda_{I_{CE}} = 0.25$. For E_{ME} and I_{CE}, we use $\alpha = 0.1$, $\beta = 0.005$, and $\gamma = 1$ in Eqn. (3) and Eqn. (5). E_t are selected within 50ms before the timestamps of I_t and processed in the voxel grid representation [43]. It takes 40,000 iterations on a batch size of two to train our CMDA. All experiments are conducted on a Tesla A100 GPU.

4.2. Datasets

DSEC Night-Semantic Dataset. To provide a benchmark for nighttime image-event semantic segmentation, we introduce the first image-event nighttime semantic segmentation dataset, *i.e.,* DSEC Night-Semantic, based on the DSEC dataset [13]. In DSEC, images and events are acquired by two different sensors which makes the two modalities not completely aligned. To obtain paired image-event data, we utilize depth data to warp from the image coordinates to the event coordinates with a resolution of 640×480. Our dataset consists of 5 nighttime sequences of Zurich City 09a-e, and includes 1,692 training samples and 150 testing samples. For each testing sample, we manually annotate them in 18 classes: Road, Sidewalk, Building, Wall, Fence, Pole, Traffic Light, Traffic Sign, Vegetation, Terrain, Sky, Person, Rider, Car, Bus, Train, Motorcycle and Bicycle. Detailed dataset

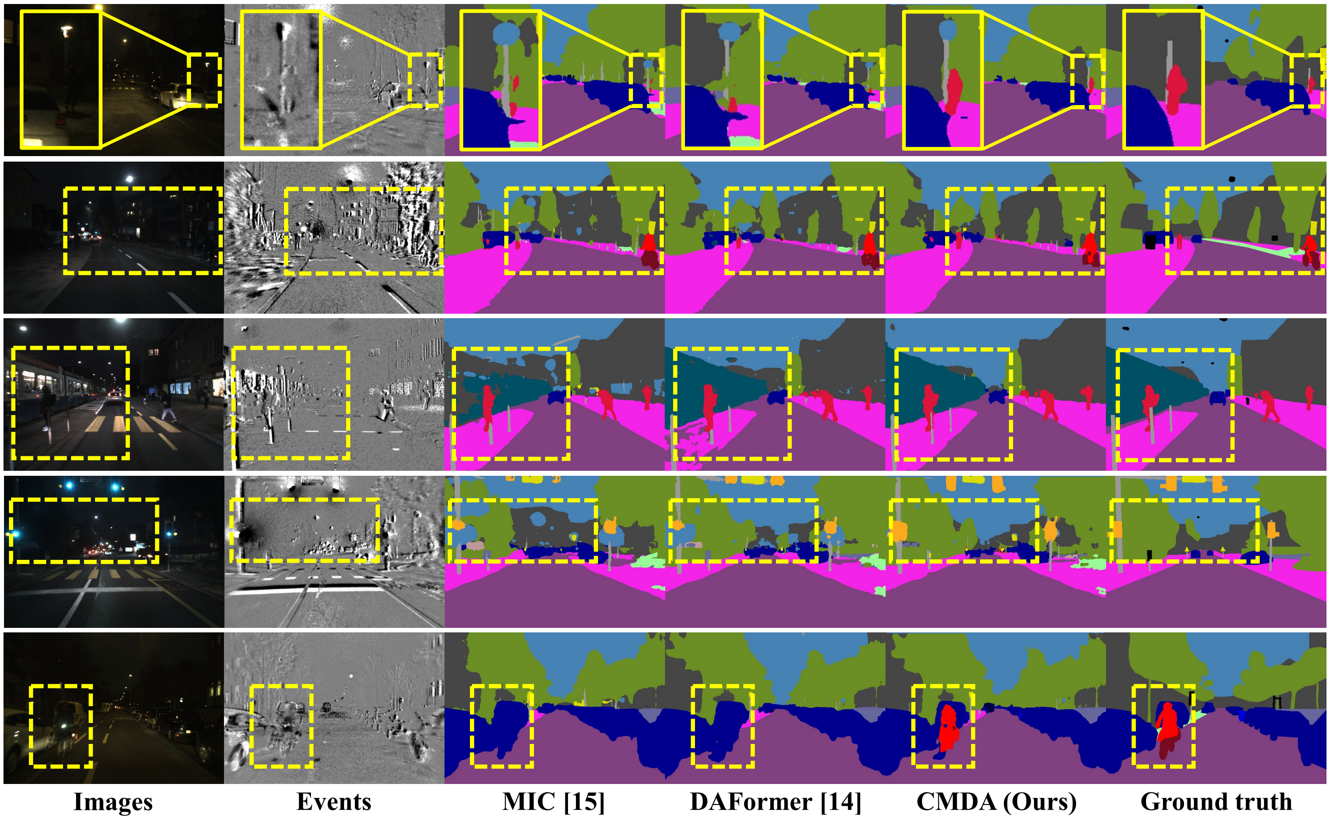

Figure 4. Qualitative semantic segmentation results generated by image-based SOTA methods MIC [15], DAFormer [14], and our proposed CMDA in the DSEC Night-Semantic dataset.

Method	Road	S.walk	Build.	Wall	Fence	Pole	Tr.L.	Tr.S.	Veg.	Terr.	Sky	Person	Rider	Car	Bus	Train	M.bike	Bike	MIoU
SePiCo† [35]	93.3	58.7	56.8	28.2	4.7	34.1	27.9	55.1	55.7	56.1	76.1	50.5	30.5	75.1	75.5	71.0	22.6	26.6	49.9
Refign† [3]	92.2	56.6	**59.2**	28.0	7.9	38.4	32.1	**60.0**	56.9	57.5	79.6	**60.3**	26.3	72.3	68.7	77.8	39.3	35.7	52.7
MIC [15]	94.0	62.1	54.2	36.3	**9.8**	37.7	29.2	48.4	**62.6**	67.2	74.5	53.1	25.5	73.0	79.7	65.7	56.0	37.4	53.7
DAFormer [14]	93.9	64.3	53.7	34.9	7.5	40.7	34.1	55.9	61.6	68.7	84.5	57.1	28.8	75.0	68.5	77.8	57.6	42.6	56.0
Baseline(I)	94.2	64.5	44.8	36.3	**9.8**	39.1	23.8	58.3	56.5	67.3	73.0	59.5	34.4	75.4	87.6	**78.8**	42.6	45.2	55.1
CMDA(E)	90.8	50.9	59.1	30.5	4.4	26.2	28.1	41.6	53.5	49.6	68.3	33.9	30.2	68.0	65.5	57.3	41.9	28.6	46.0
CMDA(I)	**94.6**	67.5	55.5	36.2	7.9	39.3	42.2	55.6	60.7	70.2	**85.4**	50.7	39.3	77.6	84.8	73.9	53.2	**45.3**	57.8
CMDA(I+E)	**94.6**	**68.3**	58.2	**37.5**	8.8	**44.0**	**45.7**	57.7	61.4	**70.4**	85.1	56.0	**45.9**	**79.2**	**87.8**	73.8	**61.6**	45.0	**60.1**

Table 2. Quantitative semantic segmentation results evaluated with MIoU (%) in our proposed DSEC Night-Semantic Dataset. (I/E/I+E) indicates the input modalities during testing. The best result is highlighted in bold. † denotes the methods utilizing additional coarsely aligned daytime images in the target domain which are not available in our dataset. We directly test their model trained on Dark Zurich [25].

split is shown in Table 1. Distribution of annotations across individual classes is provided in the supplemental material.

Dark Zurich Dataset. To thoroughly evaluate the effectiveness of our Image Content-Extractor, we conduct experiments on the image-based Dark Zurich dataset [25]. Since there is no event modality in this dataset, we exclude E along with steps 2 and 4 of Algorithm 1 during training.

4.3. Comparison of SOTA Approaches

DSEC Night-Semantic Dataset. First, we compare our proposed CMDA with previous SOTA image-based unsupervised nighttime semantic segmentation approaches, including SePiCo [35], Refign [3], MIC [15], and DAFormer [14]. The results in Table 2 and Figure 4 demonstrate the superior performance of our proposed CMDA, outperforming

Method	Road	S.walk	Build.	Wall	Fence	Pole	Tr.L.	Tr.S.	Veg.	Terr.	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	MIoU
MGCDA† [26]	80.3	49.3	66.2	7.8	11.0	41.4	38.9	39.0	64.1	18.0	55.8	52.1	**53.5**	74.7	**66.0**	0.0	37.5	29.1	22.7	42.5
DANNet† [33]	90.0	54.0	74.8	41.0	21.1	25.0	26.8	30.2	72.0	26.2	84.0	47.0	33.9	68.2	19.0	0.3	66.4	38.3	23.6	44.3
CDAda† [37]	90.5	60.6	67.9	37.0	19.3	42.9	36.4	35.3	66.9	24.4	79.8	45.4	42.9	70.8	51.7	0.0	29.7	27.7	26.2	45.0
DANIA† [34]	90.8	59.7	73.7	39.9	**26.3**	36.7	33.8	32.4	70.5	32.1	85.1	43.0	42.2	72.8	13.4	0.0	71.6	48.9	23.9	47.2
CCDistill† [11]	89.6	58.1	70.6	36.6	22.5	33.0	27.0	30.5	68.3	33.0	80.9	42.3	40.1	69.4	58.1	0.1	72.6	47.7	21.3	47.5
LoopDA† [27]	92.1	63.3	**80.3**	41.1	13.9	40.8	39.7	41.1	71.3	28.4	85.5	50.2	38.5	78.2	58.5	3.0	77.2	26.5	31.0	50.6
DAFormer [14]	93.5	65.5	73.3	39.4	19.2	53.3	44.1	44.0	59.5	**34.5**	66.6	53.4	52.7	82.1	52.7	9.4	89.3	50.5	38.5	53.8
SePiCo† [35]	93.2	68.1	73.7	32.8	16.3	54.6	**49.5**	48.1	**74.2**	31.0	**86.3**	57.9	50.9	82.4	52.2	1.3	83.8	43.9	29.8	54.2
MIC [15]	88.2	60.5	73.5	**53.5**	23.8	52.3	44.6	43.8	68.6	34.0	58.1	57.8	48.2	78.7	58.0	**13.3**	**91.2**	46.1	**42.9**	54.6
Baseline	**94.3**	**70.0**	77.4	40.8	13.8	53.3	28.9	44.7	66.4	34.1	81.4	57.1	42.7	81.3	49.6	5.0	89.4	50.5	35.8	53.5
Base.+MGCDA	93.7	68.7	76.8	40.1	26.1	**56.9**	49.0	**55.3**	37.9	30.2	20.8	**59.3**	49.6	**83.9**	28.9	4.3	85.0	**52.3**	34.1	50.2
CMDA(I)	93.4	65.6	76.0	40.9	22.4	54.8	48.5	47.6	65.7	30.2	78.1	56.8	46.9	80.8	64.2	12.9	74.7	44.5	37.0	**54.8**

Table 3. Quantitative semantic segmentation results evaluated with MIoU (%) in the image-based Dark Zurich Dataset. The best result is highlighted in bold.

Method	MIoU(E)	MIoU(I)	MIoU(I+E)
Baseline	-	55.06	-
Base. w/ I_{CE}	-	56.78	-
Base. w/ E_{ME}	45.06	53.46	55.65
CMDA	**46.02**	**57.76**	**60.05**

Table 4. Ablation of I_{CE} and E_{ME} in our CMDA.

DAFormer [14] by +4.1%. The fusion of high dynamic range event modality facilitates robust feature extracting from the scene, achieving improved nighttime semantic segmentation of 60.1%. In addition, we find that training with the event modality and testing without it is also instrumental. The performance of CMDA(I) is significantly improved compared to the baseline (+2.7%), which indicates that events can guide the network in extracting more reliable features from images at night. Qualitative results in Figure 4 demonstrate the substantial improvement in the segmentation of low-light objects and backgrounds.

Dark Zurich Dataset. In Table 3, we conduct experiments on the image-based Dark Zurich dataset to verify the effectiveness of our proposed Image Content-Extractor. First, we combine the day-to-night style transfer network of MGCDA [26] with our baseline, and style transfer on the input domain is supposed to help the self-training framework in DAFormer [14] to alleviate the domain adaptation difficulties. However, the result is degraded (-3.3%) due to the unrealistic and unreliable transferred images. In contrast, our proposed Image Content-Extractor eliminates most of the style information while preserving the content information, which surpasses the baseline by +1.3% and achieves the SOTA MIoU score of 54.8%.

4.4. Ablation Studies

Image Content-Extractor and Image Motion-Extractor are key components of the CMDA framework, bridging the gaps between domains and modalities. Table 4 provides an overview of the ablation studies of these two components. (1) The application of I_{CE} results in an improvement of the baseline performance MIoU(I) by +1.72%, demonstrating the assistance of I_{CE} for minimizing the domain shifts between the representations of daytime and nighttime images. (2) However, introducing event modality with only E_{ME} impairs the features extraction of image. MIoU(I) has a reduction of -1.6% compared to the baseline and MIoU(I+E) only has a minor improvement of +0.6%. We consider that when calculating $\mathcal{L}_t$, pseudo labels $\hat{Y}_t$ are generated by the fusion of both modalities. However, this fusion is unreliable at the beginning and hinders the initial training of the network, which in turn has a detrimental effect. (3) When employing both I_{CE} and E_{ME}, we fuse I and E/I_{CE} randomly at each training step, which alleviates the above problem. The performance is further improved to 60.05% MIoU(I+E), improving +4.99% compared to the baseline. More detailed ablation studies of the Image Motion-Extractor and Image Content-Extractor are shown below.

4.5. Image Motion-Extractor

We compare our Image Motion-Extractor with ESIM [22] and EventGAN [42] that directly generate events from two temporally adjacent images, and a straightforward approach that generates events from daytime images by a style transfer network G. Results are presented in Table 5 and Figure 5.

As demonstrated in Table 5, our proposed E_{ME} exhibits superior MIoU(E) performance compared to ESIM [22] (+2.82%) and EventGAN [42] (+1.23%), even when implemented without G. When combined with G, the proposed

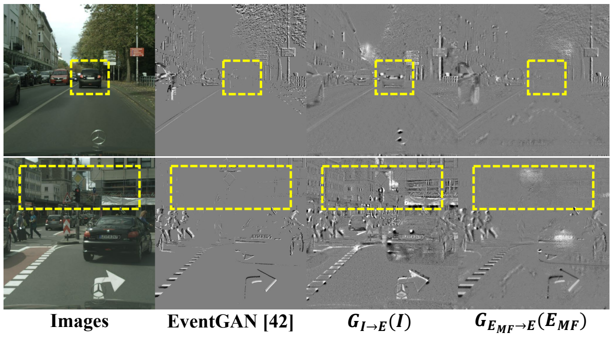

Figure 5. Comparison of different ways to generate $\hat{E}_s$. As shown in the yellow box, $\hat{E}_s$ generated from a single image $G_{I\to E}(I)$ cannot simulate motion-related regions, which has a significant distribution difference from real events. In addition, $\hat{E}_s$ from EventGAN [42] does not construct the nighttime style.

Method	MIoU(E)	MIoU(I)	MIoU(I+E)
ESIM[22] $\rightarrow E_t$	42.09	53.59	54.10 (+0.51)
$I \rightarrow E_t$	41.81	54.21	54.50 (+0.29)
$E_{ME} \rightarrow E_t$	44.91	55.47	56.63 (+1.16)
EventGAN[42] $\rightarrow E_t$	43.68	55.79	56.74 (+0.95)
$I + G \rightarrow E_t$	39.03	55.24	57.21 (+1.97)
$E_{ME} + G \rightarrow E_t$	**46.02**	**57.76**	**60.05 (+2.29)**

Table 5. Different approaches of adapting to nighttime event modality. The values in parentheses of MIoU(I+E) represent the gain compared to MIoU(I) after fusion with the event modality.

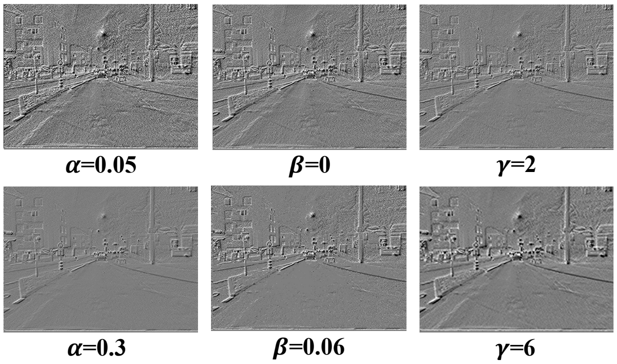

Figure 6. Visualization of nighttime I_{CE} generated with different parameters.

$E_{ME} + G$ achieves a remarkable improvement of +2.29%. It surpasses the improvement +1.97% of $I + G$ and achieves the SOTA performance of 60.05%.

Visualization of $\hat{E}_s$ is shown in Figure 5. EventGAN [42] ignores the noise of event cameras at night, and $\hat{E}_s$ generated by I depicts all edges in the scene, which fails to accurately simulate the motion-capture property of event cameras. By employing E_{ME} with G, our $\hat{E}_s$ simulates events only in the regions with the relative motion and achieves a more accurate depiction of nighttime events.

α	0.05	0.1	0.15	0.2
MIoU(I+E)	57.59	**60.05**	59.43	59.70

β	0	0.005	0.015	0.03
MIoU(I+E)	58.38	**60.05**	59.04	57.61

γ	1	2	3	4
MIoU(I+E)	**60.05**	59.40	59.28	58.57

Table 6. Analysis of α, β and γ. When adjusting one parameter, the other two parameters in the gray background remain unchanged.

4.6. Image Content-Extractor

Our Image Content-Extractor plays a key role in bridging the domain gap between daytime and nighttime images. In Figure 6, we provide a visualization of nighttime I_{CE} generated with α, β in Eq. 3 and γ in Eq. 5. α controls the lower-bound and upper-bound of $\mathrm{F}_{\mathrm{LogDiff}}(I_1, I_2)$. A large value of α narrows down the effective information in the scene, while a small value of α amplifies the proportion of noise. β aims to filter out the values less than β. A smaller β will retain more noise while a larger β will destroy the information of the scene. γ controls the shift pixels of the image relative to itself. A small value of γ can better capture scene details. Conversely, a large value of γ blurs the edges. Experiments in Table 6 demonstrate that the moderate values of α, β, and small value of γ have the optimal trade-off.

5. Conclusion

We introduce a novel framework, Cross-Modality Domain Adaptation (CMDA), for semantic segmentation on nighttime image and event modalities. Our proposed Image Motion-Extractor and Image Content-Extractor effectively bridge the gaps between modalities and domains. Notably to the best of our knowledge, our work is the first to introduce event modality into nighttime semantic segmentation. To facilitate our research, we present the DSEC Night-Semantic dataset that comprises 1,692 training samples and 150 testing samples. A comprehensive evaluation demonstrates that our CMDA achieves substantial performance improvements and effectively leverages the complementary modalities.

Acknowledgment. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB1714300, in part by the National Natural Science Foundation of China under Grant 62233005, in part by the Program of Shanghai Academic Research Leader under Grant 20XD1401300, in part by the Sino-German Center for Research Promotion under Grant M-0066, and in part by the Program of Introducing Talents of Discipline to Universities through the 111 Project under Grant B17017.

References

[1] Inigo Alonso and Ana Murillo. EV-SegNet: Semantic segmentation for event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1624–1633, 2019. 2

[2] Jonathan Binas, Daniel Neil, Shih Liu, and Tobi Delbruck. DDD17: End-to-end DAVIS driving dataset. *ArXiv:1711.01458*, 2017. 2

[3] David Brüggemann, Christos Sakaridis, Prune Truong, and Luc Van Gool. Refign: Align and refine for adaptation of semantic segmentation to adverse conditions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3174–3184, 2023. 6

[4] Lyujie Chen, Yao Xiao, Xiaming Yuan, Yiding Zhang, and Jihong Zhu. Robust autonomous landing of UAVs in non-cooperative environments based on comprehensive terrain understanding. *Science China Information Sciences*, 65(11):212202, 2022. 1

[5] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 1

[6] 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. 2

[7] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *International Conference on Intelligent Transportation Systems*, pages 3819–3824. IEEE, 2018. 3

[8] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Conference on Robot Learning*, pages 1–16. PMLR, 2017. 2

[9] Thomas Finateu, Atsumi Niwa, Daniel Matolin, Koya Tsuchimoto, Andrea Mascheroni, Etienne Reynaud, Pooria Mostafalu, Frederick Brady, Ludovic Chotard, Florian LeGoff, Hirotsugu Takahashi, Hayato Wakabayashi, Yusuke Oike, and Christoph Posch. A 1280×720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86μm pixels, 1.066 geps readout, programmable event-rate controller and compressive data-formatting pipeline. In *IEEE International Solid-State Circuits Conference*, pages 112–114, 2020. 1

[10] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2020. 2, 3

[11] Huan Gao, Jichang Guo, Guoli Wang, and Qian Zhang. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9913–9923, 2022. 1, 2, 3, 4, 7

[12] Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics and Automation Letters*, 6(2):2822–2829, 2021. 2

[13] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. 2, 5

[14] 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, 4, 5, 6, 7

[15] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. MIC: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11721–11732, 2023. 3, 6, 7

[16] Seunghun Lee, Sunghyun Cho, and Sunghoon Im. DRANet: Disentangling representation and adaptation networks for unsupervised cross-domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15252–15261, 2021. 2

[17] 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

[18] Chuming Lin, Bo Yan, and Weimin Tan. Foreground detection in surveillance video with fully convolutional semantic network. In *IEEE International Conference on Image Processing*, pages 4118–4122. IEEE, 2018. 1

[19] Chenxin Liu, Jiahu Qin, Shuai Wang, Lei Yu, and Yaonan Wang. Accurate RGB-D SLAM in dynamic environments based on dynamic visual feature removal. *Science China Information Sciences*, 65(10):202206, 2022. 1

[20] Marin Orsic, Ivan Kreso, Petra Bevandic, and Sinisa Segvic. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12607–12616, 2019. 1

[21] Hong Qiao, Shanlin Zhong, Ziyu Chen, and Hongze Wang. Improving performance of robots using human-inspired approaches: a survey. *Science China Information Sciences*, 65(12):221201, 2022. 1

[22] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: An open event camera simulator. In *Conference on Robot Learning*, pages 969–982. PMLR, 2018. 7, 8

[23] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):1964–1980, 2019. 2

[24] Eduardo Romera, Luis M Bergasa, Kailun Yang, Jose M Alvarez, and Rafael Barea. Bridging the day and night domain gap for semantic segmentation. In *IEEE Intelligent Vehicles Symposium*, pages 1312–1318. IEEE, 2019. 3

[25] 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. 2, 3, 6

[26] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3139–3153, 2020. 3, 7

[27] Fengyi Shen, Zador Pataki, Akhil Gurram, Ziyuan Liu, He Wang, and Alois Knoll. LoopDA: Constructing self-loops to adapt nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3256–3266, 2023. 7

[28] Mennatullah Siam, Mostafa Gamal, Moemen Abdel-Razek, Senthil Yogamani, Martin Jagersand, and Hong Zhang. A comparative study of real-time semantic segmentation for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 587–597, 2018. 1

[29] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021. 1

[30] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. ESS: Learning event-based semantic segmentation from still images. *European Conference on Computer Vision*, 2022. 2

[31] Lin Wang, Yujeong Chae, and Kuk Yoon. Dual transfer learning for event-based end-task prediction via pluggable event to image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2135–2145, 2021. 2

[32] Lin Wang, Yujeong Chae, Sung Yoon, Tae Kim, and Kuk Yoon. EvDistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 608–619, 2021. 2

[33] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. DANNet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15769–15778, 2021. 1, 3, 7

[34] Xinyi Wu, Zhenyao Wu, Lili Ju, and Song Wang. A one-stage domain adaptation network with image alignment for unsupervised nighttime semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):58–72, 2021. 1, 3, 7

[35] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. SePiCo: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 6, 7

[36] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1, 5

[37] Qi Xu, Yinan Ma, Jing Wu, Chengnian Long, and Xiaolin Huang. CDAda: A curriculum domain adaptation for nighttime semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2962–2971, 2021. 1, 2, 3, 4, 7

[38] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023. 2

[39] Jiaming Zhang, Kailun Yang, and Rainer Stiefelhagen. ISSAFE: Improving semantic segmentation in accidents by fusing event-based data. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1132–1139, 2021. 2

[40] Chaoqiang Zhao, Yang Tang, and Qiyu Sun. Unsupervised monocular depth estimation in highly complex environments. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(5):1237–1246, 2022. 2, 4

[41] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2582–2593, 2022. 1

[42] Alex Zihao Zhu, Ziyun Wang, Kaung Khant, and Kostas Daniilidis. EventGAN: Leveraging large scale image datasets for event cameras. In *IEEE International Conference on Computational Photography*, pages 1–11. IEEE, 2021. 2, 7, 8

[43] 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. 5

[44] 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. 2, 4