ICDA: Illumination-Coupled Domain Adaptation Framework for Unsupervised Nighttime Semantic Segmentation

Chenghao Dong , Xuejing Kang* , Anlong Ming
School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications
{chdong, kangxuejing, mal}@bupt.edu.cn

Abstract

The performance of nighttime semantic segmentation has been significantly improved thanks to recent unsupervised methods. However, these methods still suffer from complex domain gaps, i.e., the challenging illumination gap and the inherent dataset gap. In this paper, we propose the illumination-coupled domain adaptation framework(ICDA) to effectively avoid the illumination gap and mitigate the dataset gap by coupling daytime and nighttime images as a whole with semantic relevance. Specifically, we first design a new composite enhancement method(CEM) that considers not only illumination but also spatial consistency to construct the source and target domain pairs, which provides the basic adaptation unit for our ICDA. Next, to avoid the illumination gap, we devise the Deformable Attention Relevance(DAR) module to capture the semantic relevance inside each domain pair, which can couple the daytime and nighttime images at the feature level and adaptively guide the predictions of nighttime images. Besides, to mitigate the dataset gap and acquire domain-invariant semantic relevance, we propose the Prototype-based Class Alignment(PCA) module, which improves the usage of category information and performs fine-grained alignment. Extensive experiments show that our method reduces the complex domain gaps and achieves state-of-the-art performance for nighttime semantic segmentation. Our code is available at https://github.com/chenghaoDong666/ICDA.

1 Introduction

Nighttime semantic segmentation aims to label each pixel of an image in the nighttime with a corresponding class, which is as essential as daytime for safety-critical tasks such as autonomous driving[Geiger *et al.*, 2012] but more difficult due to scarce labeled datasets and poor illumination. Limited by the difficulty of building high-quality pixel-level annotations, unsupervised domain adaptation(UDA) is widely adopted in nighttime semantic segmentation to utilize the knowledge from daytime images. However, simply adopting the general UDA models[Tsai *et al.*, 2018; Vu *et al.*, 2019; Li *et al.*, 2019; Toldo *et al.*, 2021] increases little performance due to the complex domain gaps, i.e., the challenging illumination gap and the inherent dataset gap. Compared to the dataset gap which is usual in UDA tasks[Toldo *et al.*, 2020], the illumination gap is generated by poor illumination in nighttime images and is more difficult and critical.

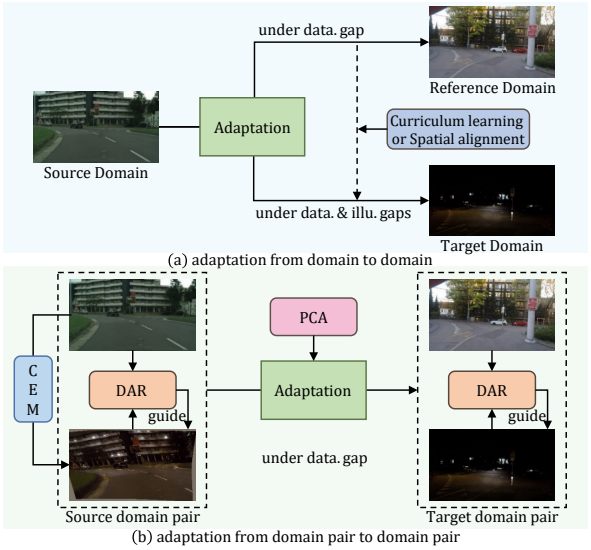

Figure 1: The differences between our ICDA and existing methods. (a) The existing methods perform adaptation from domain to domain with a separate view by curriculum learning or spatial alignment. (b) Our ICDA performs adaptation from domain pair to domain pair with a united view by coupling daytime and nighttime images.

There have already been some works focusing on solving the performance drop caused by the complex domain gaps as Figure 1 (a). Some works[Dai and Van Gool, 2018; Sakaridis *et al.*, 2019; Sakaridis *et al.*, 2020; Xu *et al.*, 2021] adopt curriculum learning which decomposes the complex gaps into the dataset gap and several smaller illumination gaps, e.g. day-twilight and twilight-night, to achieve smoother adaptation. Other works[Wu *et al.*, 2021a; Wu *et al.*, 2021b; Bruggemann *et al.*, 2023; Gao *et al.*, 2022] utilize the pseudo

*Corresponding Author

supervision from the adaptation under the dataset gap to the adaptation under the dataset and illumination gaps, which is constructed by the spatial alignment, to provide additional supervision information and facilitate the knowledge transfer. However, these methods all treat the daytime and nighttime images with a separate view when performing adaptation and thus have to face knowledge transfer under difficult illumination gap, which limits their performance.

In this paper, we propose a novel illumination-coupled domain adaptation(ICDA) framework for nighttime semantic segmentation, which treats the corresponding daytime and nighttime images with a united view and couples them as a whole through semantic relevance, thereby largely simplifying the complex domain gaps, as Figure 1 (b). Specifically, we first propose the Composite Enhance Method(CEM), which maintains the consistency of illumination and spatial differences during image enhancement to construct the source and target domain pairs. Our CEM provides the basic adaptation unit since it ensures the illumination gap exists only inside each domain pair while the dataset gap exists only between the two domain pairs. Next, we propose the Deformable Attention Relevance(DAR) module, which adopts the cross-domain deformable attention to fully capture semantic relevance inside each domain pair. With it, each domain pair is coupled at the feature level, and the predictions of nighttime images can be adaptively guided, thus avoiding the illumination gap. Moreover, to mitigate the dataset gap and facilitate the transfer of semantic relevance, we propose the Prototype-based Class Alignment(PCA) module, which performs fine-grained class alignment by controlling the distances between features and prototypes. Finally, we evaluate our ICDA on Dark Zurich[Sakaridis *et al.*, 2019] and BDD100k-night[Yu *et al.*, 2020; Sakaridis *et al.*, 2020] datasets. Our main contributions are summarized as follows:

1. The proposed CEM considers not only illumination but also spatial consistency to construct the source and target domain pairs, which provides the basic adaptation unit.

2. The proposed DAR adopts cross-domain deformable attention to capture the semantic relevance, which acts as the link to coupling daytime and nighttime images and adaptively guides the predictions of nighttime images, thus avoiding the illumination gap.

3. The proposed PCA utilizes category information and performs fine-grained class alignment which mitigates the dataset gap and acquires the domain-invariant semantic relevance between the two coupled domain pairs.

4. Extensive experiments verify that our ICDA achieves a new state-of-the-art performance on nighttime benchmarks.

2 Related Work

2.1 Domain Adaptation for Semantic Segmentation

The goal of domain adaptation for unsupervised semantic segmentation is to transfer the knowledge learned from the source to the target domain, despite the inconsistent data distributions between them. Some works adopt adversarial learning to diminish the distribution shift at image-level[Hoffman *et al.*, 2018; Yang and Soatto, 2020], feature level[Pan *et al.*, 2020; Huang *et al.*, 2020] or output level[Tsai *et al.*, 2018; Vu *et al.*, 2019]. However, these works are less stable since they only align the distribution from the holistic view while ignoring category information. Therefore, other works[Luo *et al.*, 2019; Wang *et al.*, 2020; Ma *et al.*, 2021; Zhang *et al.*, 2021; Jiang *et al.*, 2022] perform the feature alignment in a class-wise manner to achieve a more fine-grained domain adaptation. Besides, a line of works[Wang *et al.*, 2021; Hoyer *et al.*, 2022] adopt the self-training strategy which assigns pseudo labels for unlabeled target data to enrich the training data. Although these methods achieve competitive performance, they are all designed for daytime scenes, and their performance drops dramatically when faced with the complex domain gaps of nighttime scenes. In this paper, our ICDA adopts class-level alignment and self-training strategy but focuses on the complex domain gaps between daytime and nighttime images.

2.2 Nighttime Semantic Segmentation

An early idea to solve the nighttime semantic segmentation is to adopt curriculum learning thus achieving smoother adaptation. For example, some works[Dai and Van Gool, 2018; Sakaridis *et al.*, 2019; Sakaridis *et al.*, 2020; Xu *et al.*, 2021] leverage the intermediate twilight domain to decompose the complex gaps into the dataset gap and several smaller illumination gaps, e.g. day-twilight and twilight-night, thus reducing the difficulty by gradual adaptation. Although the difficulty of each stage adaptation is reduced, the multi-stage training needs additional data and is time-consuming. Another line of works instead utilize the pseudo supervision from the adaptation under the dataset gap to the adaptation under the dataset and illumination gaps, which is constructed by the static loss[Wu *et al.*, 2021a; Wu *et al.*, 2021b; Gao *et al.*, 2022] or spatial alignment[Wu *et al.*, 2021b; Bruggemann *et al.*, 2023], to provide additional supervision information and improve the performance. However, the performance of spatial alignment is restricted by dynamic and small static object regions. More importantly, the above two types of methods all treat the daytime and nighttime images with a separate view when performing the adaptation and thus have to perform the knowledge transfer under the difficult illumination gap. In this paper, our ICDA instead addresses the complex domain gaps with a united view by coupling daytime and nighttime images as a whole, which successfully avoids the illumination gap and mitigates the dataset gap.

3 Method

In this section, we first introduce our ICDA framework which largely simplifies the complex domain gaps in general. We then introduce the CEM which constructs the source and target domain pairs. Subsequently, we introduce the DAR and PCA which captures the semantic relevance and facilitates the transfer of it in class-level separately. Finally, the overall objective functions are presented.

3.1 Overview

The whole framework of our ICDA is depicted as Figure 2, which consists of three parts: the CEM part, the input part,

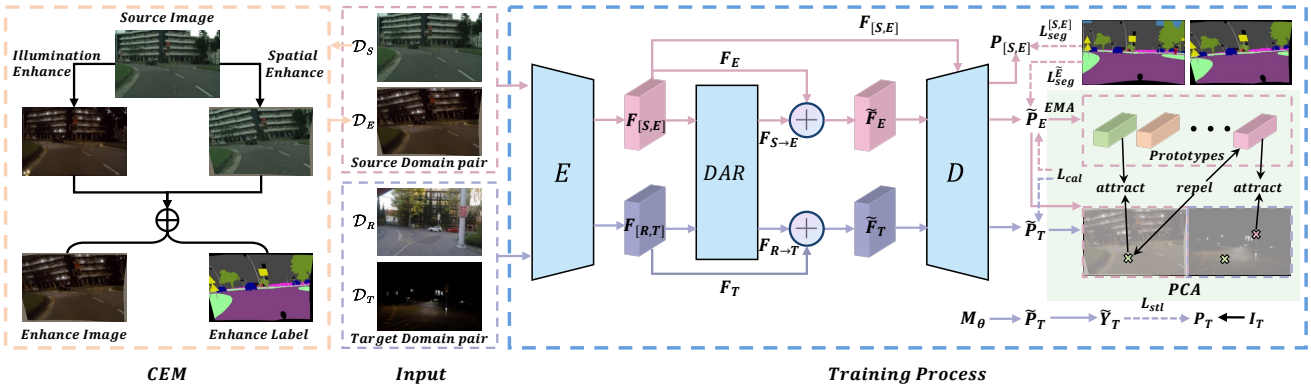

Figure 2: The overall framework of our ICDA.

and the training process part. The CEM part generates the enhance domain $\mathcal{D}_E$ from the source domain $\mathcal{D}_S$ to constitute the source domain pair, which is consistent with the target domain pair in illumination and spatial differences. In the input part, our method takes the source domain pair and target domain pair as the input unit, which contains the source domain $\mathcal{D}_S$, enhance domain $\mathcal{D}_E$ and reference domain $\mathcal{D}_R$, target domain $\mathcal{D}_T$, respectively. Only the source domain pair has semantic labels. In the training process part, our ICDA adopts the DAFormer[Hoyer *et al.*, 2022] as the backbone, which extracts and decodes the features by encoder E and decoder D, and performs the online self-training with the Mean teacher[Tarvainen and Valpola, 2017] strategy to generate convincing pseudo labels $\tilde{Y}_T$ using teacher model M_θ and improve the robustness of the model. Based on this, the proposed DAR captures the semantic relevance features $F_{S\to E}/F_{R\to T}$ from domain pair features $F_{[S,E]}/F_{[R,T]}$ then uses them to guide the nighttime features F_E/F_T and gets the remedied features $\tilde{F}_E/\tilde{F}_T$. And the PCA module mitigates the dataset gap and captures the domain-invariant semantic relevance by aligning the remedied prediction maps $\tilde{P}_E$ and $\tilde{P}_T$ at the class level. The CEM, DAR, and PCA will be introduced in detail next.

3.2 Composite Enhancement Method

Apart from the illumination consistency focused on by the existing methods[Sakaridis *et al.*, 2019; Sakaridis *et al.*, 2020; Xu *et al.*, 2021; Wu *et al.*, 2021a; Gao *et al.*, 2022] in image enhancement, we also take the spatial consistency into account. Without spatial consistency, the semantic relevance captured from the source domain pair is not robust to the spatial differences when transferred to the target domain pair, since the corresponding reference and target domain images are captured from different viewpoints of the same scene. Therefore, our CEM composes two types of enhancement: illumination and spatial enhancement.

Illumination Enhancement. For illumination enhancement, we adopt the CycleGAN[Zhu *et al.*, 2017] which is widely used for unsupervised image enhancement to generate the illumination changes from daytime to nighttime.

Spatial Enhancement. For spatial enhancement, similar to [Truong *et al.*, 2021; Rocco *et al.*, 2017], we construct spatial changes by randomly sampling three transformations: homography, Thin-plate Spline(TPS), and affine-TPS. The homography and TPS generate perspective changes and smooth deformation separately while the affine-TPS further extends the TPS using affine transformations to bring larger scale, angle, and shape changes. Note that different from [Truong *et al.*, 2021; Rocco *et al.*, 2017], our spatial enhancement is not only applied to images but also semantic labels so that the source domain pair can perform supervised learning.

Benefiting from our CEM, the illumination and spatial differences inside the source domain pair are consistent with those inside the target domain pair. In this way, our CEM ensures that the illumination gap exists only inside each domain pair while the dataset gap exists only between the two domain pairs, which provides the basic adaptation unit and is necessary to perform the DAR and PCA.

3.3 Deformable Attention Relevance

In this section, we propose the DAR module to capture the semantic relevance inside each domain pair, which acts as the link to couple daytime and nighttime images as a whole and can adaptively guide the predictions of nighttime images. As shown in Figure 3, our DAR constructs a lightweight transformer with its basic block $\times 2$, which contains the local window attention(LWA) and cross-domain deformable attention(CDDA). The LWA aggregates local features and is only applied to nighttime features to achieve the trade-off between performance and efficiency. And the CDDA captures cross-domain semantic relevance against the illumination and spatial differences inside each domain pair, which is the core of our DAR and will be described in detail below.

Our CDDA mainly contains the *deformed feature resampling* and the *cross-domain multi-head attention*. For the *deformed feature resampling*, we first get the fusion of the daytime and nighttime features and then feed it to our offset generator. Different from DCN[Dai *et al.*, 2017] that uses convolution to predict local spatial offset, our offset generator consists of a basic transformer block to acquire a global receptive field and a depthwise separable convolution to downsample the channels, thus getting the global spatial offset. We then acquire the deformed sampling points by adding the uniform grid of reference points similar to the DAT[Xia *et al.*, 2022]

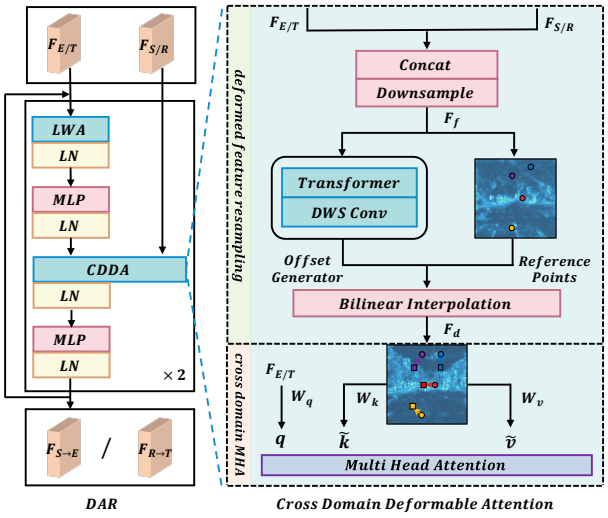

Figure 3: The deformable attention relevance module DAR.

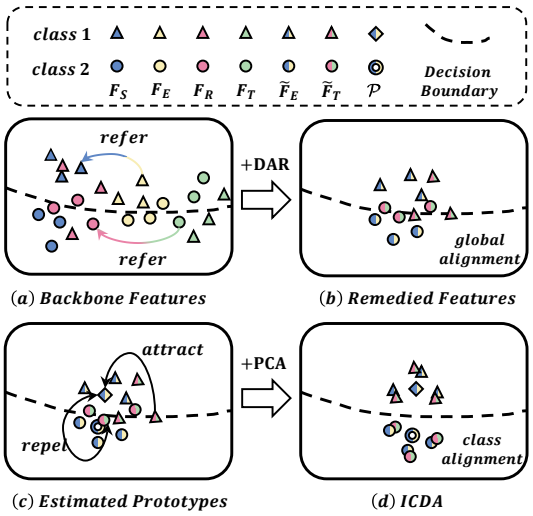

Figure 4: Illustration of the global and class alignment effects of DAR and PCA.

with our global spatial offset. With the deformed sampling points, our CDDA gets the deformed feature by using bilinear interpolation to resample the fused feature. The process can be formulated as follows:

$$F_f = \mathbb{D}(\mathbb{C}(F_{E/T}, F_{S/R})), F_d = \phi(F_f, p + \triangle p) \quad (1)$$

where $F_{E/T}, F_{S/R} \in \mathbb{R}^{H\times W\times C}$ denotes the nighttime and daytime features, respectively. The fused feature $F_f \in \mathbb{R}^{H\times W\times C}$ is acquired by the concatenation $\mathbb{C}$ and downsample convolution $\mathbb{D}$. The $\triangle p, p \in \mathbb{R}^{H\times W\times 2}$ are the global spatial offset and the reference points. With the guidance of p and $\triangle p$, the fused feature F_f is resampled towards important regions using the bilinear interpolation ϕ and the deformed feature $F_d \in \mathbb{R}^{H\times W\times C}$ is acquired.

As for the *cross-domain multi-head attention*, different from the traditional attention whose query, key, and value are the same, our CDDA instead uses the nighttime feature as the query and the deformed feature as the deformed key and value in the multi-head attention as below:

$$q = F_{E/T} \times W_q, \tilde{k} = F_d \times W_k, \tilde{v} = F_d \times W_v \quad (2)$$

where W_q, W_k, W_v are projection matrices, and $q, \tilde{k}, \tilde{v}$ are the query, deformed key, and deformed value. Hence, the features of different domains can interact with each other and we finally get the semantic relevance feature $F_{S\to E}/F_{R\to T} \in \mathbb{R}^{H\times W\times C}$, which couples the daytime and nighttime images at the feature level.

To sum up, compared to general attention, our CDDA has two main advantages: deformable and cross-domain. (1)**Deformable**. Our CDDA first extends the DCN with global spatial offset and feature resampling to model the geometric transformation caused by viewpoint shift and adaptively aggregate the features under the guidance of the important regions inside each domain pair in a global range. (2)**Cross-domain**. Based on the deformed feature which is acquired by resampling the fusion of daytime and nighttime features, our CDDA then adopts cross-domain attention to connect the nighttime feature with the deformed feature, thus getting cross-domain semantic relevance. By transferring the learned semantic relevance from the source domain pair to the target domain pair against the dataset gap, the difficult illumination gap can be avoided since the prediction of target images can be guided by the semantic relevance without the need for additional adaptation.

As for the reason that semantic relevance can remedy the prediction of nighttime images, we use Figure 4 to clarify it. According to the illumination conditions and prediction difficulty, the source, reference, enhance and target domain features F_S, F_R, F_E, F_T lie in the discriminant space from left to right as Figure 4 (a). Benefiting from the semantic relevance, the nighttime features F_E/F_T can refer to the daytime features F_S/F_R to remedy itself, and the remedied feature distributions are globally aligned in the discriminant space as Figure 4 (b), which improves the model's performance.

3.4 Prototype-Based Class Alignment

With the two semantic-coupled domain pairs, the only obstacle to knowledge transfer is the inherent dataset gap. To further mitigate the dataset gap and capture domain-invariant semantic relevance, we propose the PCA to make use of the category information and perform the adaptation in a fine-grained way. As depicted in Figure 2, we first build the prototypes $\mathcal{P}$ using the exponential moving average which can be calculated as:

$$\hat{\mathcal{P}}_{c,t} = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W} \tilde{P}_E^{t,i,j} \mathbb{1}[Y_E^{t,i,j} = c]}{\sum_{i=1}^{H}\sum_{j=1}^{W} \mathbb{1}[Y_E^{t,i,j} = c]}$$
$$\mathcal{P}_{c,t+1} = \xi \mathcal{P}_{c,t} + (1-\xi)\hat{\mathcal{P}}_{c,t} \quad (3)$$

where $\xi \in [0, 1]$ is a momentum coefficient, $\tilde{P}_E^{t,i,j}, Y_E^{t,i,j}$ denote the remedied predictions and the labels of enhance domain images respectively, t is the current iteration, i, j denote the index of the height H and width W, c is the index of category number C, $\mathbb{1}$ is the indicator function.

After getting the prototypes $\mathcal{P}$, we perform class alignment not only between the prototypes $\mathcal{P}$ and the remedied enhance domain predictions $\tilde{P}_E$, but also between the prototypes $\mathcal{P}$ and the remedied target domain predictions $\tilde{P}_T$. The former facilitates the correctness of semantic relevance in the source domain pair, and the latter facilitates its transfer at the class level. We adopt the pixel contrastive loss proposed in [Jiang *et al.*, 2022] to perform the class-centered distribution alignment for adaptation as below:

$$\begin{aligned}\mathcal{L}_{cal} = &-\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{C}\mathbb{O}(\tilde{P}_E^{i,j,c})\log \mathcal{S}_{E\to E}^{i,j,c} \\ &-\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{C}\mathbb{O}(\tilde{P}_T^{i,j,c})\log \mathcal{S}_{T\to E}^{i,j,c}\end{aligned} \tag{4}$$

where $\mathbb{O}$ denotes one-hot encoding. $S_{E\to E}^{i,j,c}$ and $S_{T\to E}^{i,j,c}$ denote the similarity of features $\tilde{P}_E^{i,j,c}$, $\tilde{P}_T^{i,j,c}$ with prototypes $\mathcal{P}^c$ as below:

$$S_{x\to E}^{i,j,c} = \frac{\exp\left(\mathcal{P}_c \cdot \tilde{P}_x^{i,j}/\tau\right)}{\sum_{c=1}^{C}\exp\left(\mathcal{P}_c \cdot \tilde{P}_x^{i,j}/\tau\right)} \tag{5}$$

where τ is the temperature.

With the $\mathcal{L}_{cal}$, our PCA forces the features to be close to the prototypes belonging to the same category while staying away from the prototypes belonging to a different category. Through the backpropagation, the remedied features of enhance and target domain images are further aligned at the class level as shown in Figure 4 (c) and (d). And the domain-invariant semantic relevance can be captured, which can provide better guidance for the prediction of target domain images, thus further mitigating the dataset gap.

3.5 Objective Functions

The overall loss of our ICDA can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{seg}^{[S,E]} + \mathcal{L}_{seg}^{\tilde{E}} + \mathcal{L}_{cal} + \mathcal{L}_{stl} \tag{6}$$

The $\mathcal{L}_{seg}^{[S,E]}$ and $\mathcal{L}_{seg}^{\tilde{E}}$ are both the cross-entropy losses, where $\mathcal{L}_{seg}^{[S,E]}$ extract the semantic knowledge of the source domain pair, and the $\mathcal{L}_{seg}^{\tilde{E}}$ ensures the correctness of the remedied enhance domain features. The class alignment loss $\mathcal{L}_{cal}$ is as mentioned before, and the self-training loss $\mathcal{L}_{stl}$ is the same as the backbone DAFormer which exploits unlabeled target data via pseudo labels. With the help of these losses, our ICDA efficiently transfers the knowledge from source to target domain pair by coupling the daytime and nighttime images to avoid the illumination gap and performing the class-level alignment to mitigate the dataset gap, thus improving the performance of nighttime semantic segmentation.

4 Experiments

4.1 Datasets

CityScapes[Cordts *et al.*, 2016] is a large dataset of urban street scenes with pixel-level annotations of 19 semantic categories. Both the images and labels of it are at a resolution of 2,048 × 1,024. During the training, we only use its training set which contains 2,975 images as the source domain.

Dark Zurich[Sakaridis *et al.*, 2019] is the mainly used unsupervised nighttime semantic segmentation dataset with a resolution of 1,920 × 1,080. During the training process, we only use the coarsely aligned 2,416 day-night image pairs to be the target domain pair, which are all unlabeled. The dataset also contains another 201 annotated nighttime images, including 50 images for validation and 151 for testing. Since the test set is not publicly available, we submit the segmentation results to the online evaluation website to get the performance of the test set. And we show the qualitative comparison and perform the ablation study on the validation set.

BDD100k-night[Sakaridis *et al.*, 2020; Yu *et al.*, 2020]. To verify the generalization of our model, we directly use our model trained from CityScapes to Dark Zurich-N to predict the test set of BDD100k-night without additional training. The BDD100k-night contains 87 images with a resolution of 1,280 × 720 and has the same semantic labels as CityScapes.

4.2 Implementation

Network. We adopt the DAFormer[Hoyer *et al.*, 2022] as our backbone, whose encoder is MiT-B5[Xie *et al.*, 2021] pretrained on the ImageNet-1k. When backpropagating through the DAR, the parameters of the encoder and decoder are frozen to avoid overfitting. For the self-training, we use the teacher model to produce the pseudo label with the momentum of EMA set to 0.999 and the threshold of the pseudo label set to 0.968. At the inference stage, our ICDA only takes the nighttime images as input and adopts the encoder E and decoder D to output the final prediction results, improving the prediction accuracy without additional inference time.

Training details. For all experiments, we use the mean of category-wise intersection-over-union (mIoU) as the evaluation metric. The whole framework is implemented using PyTorch on a single RTX 3080-Ti GPU. We use the AdamW[Loshchilov and Hutter, 2019] as the optimizer with a weight decay of 0.01. The base learning rate is 6×10^{-5} for the encoder, DAR and 6×10^{-4} for the decoder. We use the linear learning rate warmup strategy with $t_{warm} = 1.5$k and $t_{total} = 40$k. After warmup iterations, the learning rate is decreased using the poly policy with a power of 0.9. The batch size is set to 2 for each domain. Following refign[Bruggemann *et al.*, 2023], we apply random cropping with a crop size of 512 for the source domain pair, and a crop size of 960 first, then 512 for the target domain pair.

4.3 Comparison with State-of-the-art Methods

Comparison on Dark Zurich. We compare our ICDA with four types of models including the baseline models trained on CityScapes only, the general adaptation models trained from CityScapes to Dark Zurich-N, and the state-of-the-art models based on CNN and transformer which are customized for nighttime scenes on the Dark Zurich test set. For customized models based on CNN, they all adopt one of RefineNet, Deeplab-v2, or PSPNet as their backbone which uses the ResNet101[He *et al.*, 2016], while all the transformer-based models use DAFormer as the backbone.

Method	road	side.	buil.	wall	fence	pole	light	sign	veg.	terr.	sky	pers.	rider	car	truck	bus	train	motor.	bicy.	mIoU
DeepLab-v2[Chen *et al.*, 2017]	79.0	21.8	53.0	13.3	11.2	22.5	20.2	22.1	43.5	10.4	18.0	37.4	33.8	64.1	6.4	0.0	52.3	30.4	7.4	28.8
RefineNet[Lin *et al.*, 2017]	68.8	23.2	46.8	20.8	12.6	29.8	30.4	26.9	43.1	14.3	0.3	36.9	49.7	63.6	6.8	0.2	24.0	33.6	9.3	28.5
PSPNet[Zhao *et al.*, 2017]	78.2	19.0	51.2	15.5	10.6	30.3	28.9	22.0	56.7	13.3	20.8	38.2	21.8	52.1	1.6	0.0	53.2	23.2	10.7	28.8
AdaptSegNet[Tsai *et al.*, 2018]	86.1	44.2	55.1	22.2	4.8	21.1	5.6	16.7	37.2	8.4	1.2	35.9	26.7	68.2	45.1	0.0	50.1	33.9	15.6	30.4
ADVENT[Vu *et al.*, 2019]	85.8	37.9	55.5	27.7	14.5	23.1	14.0	21.1	32.1	8.7	2.0	39.9	16.6	64.0	13.8	0.0	58.8	28.5	20.7	29.7
BDL[Li *et al.*, 2019]	85.3	41.1	61.9	32.7	17.4	20.6	11.4	21.3	29.4	8.9	1.1	37.4	22.1	63.2	28.2	0.0	47.7	39.4	15.7	30.8
UDAclustering[Toldo *et al.*, 2021]	85.5	40.9	59.2	31.2	19.5	24.0	29.9	29.4	30.6	11.2	18.4	39.1	49.7	61.5	34.9	0.0	25.8	23.2	19.0	33.3
DMAda[Dai and Van Gool, 2018]	75.5	29.1	48.6	21.3	14.3	34.3	36.8	29.9	49.4	13.8	0.4	43.3	50.2	69.4	18.4	0.0	27.6	34.9	11.9	32.1
GCMA[Sakaridis *et al.*, 2019]	81.7	46.9	58.8	22.0	20.0	41.2	40.5	41.6	64.8	31.0	32.1	53.5	47.5	75.5	39.2	0.0	49.6	30.7	21.0	42.0
MGCDA[Sakaridis *et al.*, 2020]	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
CDAda[Xu *et al.*, 2021]	91.5	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
DANNet[Wu *et al.*, 2021a]	90.4	60.1	71.0	33.6	22.9	30.6	34.3	33.7	70.5	31.8	80.2	45.7	41.6	67.4	16.8	0.0	73.0	31.6	22.9	45.2
DANIA[Wu *et al.*, 2021b]	91.5	62.7	73.9	**39.9**	25.7	36.5	35.7	36.2	71.4	35.3	82.2	48.0	44.9	73.7	11.3	0.1	64.3	36.7	22.7	47.0
CCDistill[Gao *et al.*, 2022]	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
DAFormer[Hoyer *et al.*, 2022]	92.0	63.0	67.2	28.9	13.1	44.0	42.0	42.3	70.7	28.2	83.6	51.1	39.1	76.4	31.7	0.0	78.3	43.9	26.5	48.5
SePiCo[Xie *et al.*, 2022]	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
Refign[Bruggemann *et al.*, 2023]	91.8	65.0	**80.9**	37.9	25.8	56.2	45.2	51.0	**78.7**	31.0	**88.9**	58.8	52.9	77.8	51.8	6.1	90.8	40.2	**37.1**	56.2
ICDA(ours)	**93.3**	66.5	76.7	38.9	**26.9**	**56.3**	**54.7**	**52.8**	71.0	**35.8**	84.2	**58.8**	51.7	**84.0**	56.3	**20.6**	**91.7**	**51.1**	35.2	**58.2**

Table 1: Comparison with the state-of-the-art approaches on the Dark Zurich-test set.

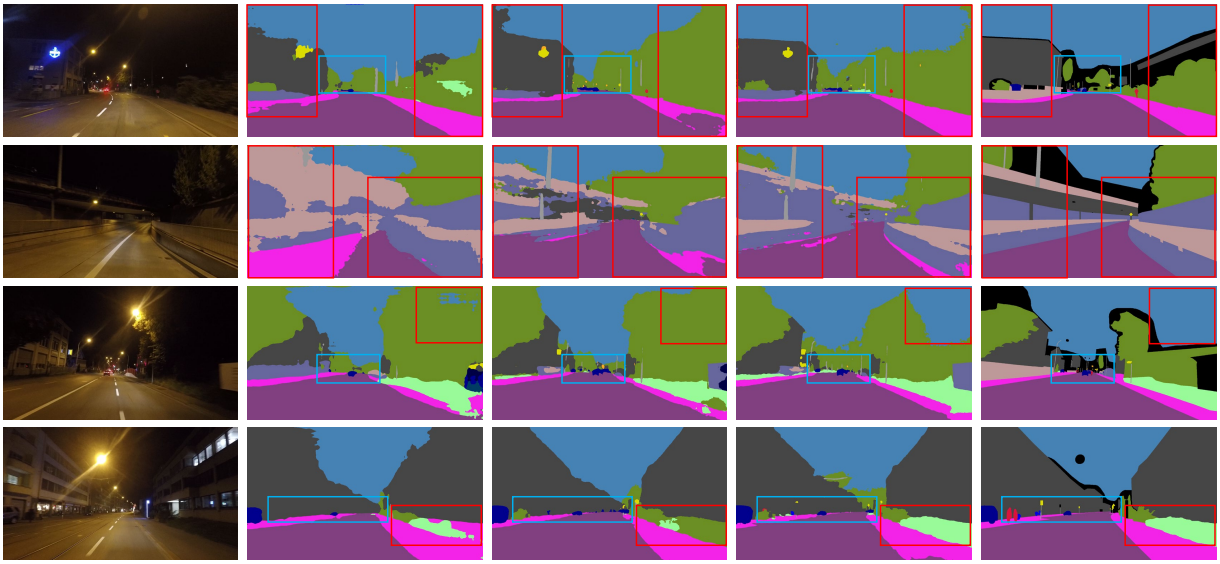

Figure 5: The qualitative comparison between our approach and some existing state-of-the-art methods on the Dark Zurich-val set.

As shown in Table 1, our method achieves the best performance(58.2%) among all of the methods, with around 2% increase compared to the previous highest score(56.2%) acquired by Refign[Bruggemann *et al.*, 2023]. For each category, our method is optimal in 11 classes, suboptimal in 4 classes, and the rest 4 classes also have competitive performances, which shows that our ICDA successfully simplifies the complex domain gaps and facilitates the knowledge transfer. Moreover, we can observe that our method achieves the best results not only in large static classes such as road, fence, and terrain but also in small static and dynamic classes such as pole, traffic light, and car. This is because our method uses cross-domain semantic relevance to adaptively guide the final predictions, so it treats each class of the 19 classes equally, unlike methods based on spatial alignment[Wu *et al.*, 2021b; Bruggemann *et al.*, 2023] which are more suitable for large static classes than the small static and dynamic classes. We also observe that our method significantly outperforms other methods on classes with relatively few occurrences, such as bus and train, which once more shows that our model makes better use of knowledge from the whole domain pair to perform the global alignment and the learned prototypes help to remedy the corresponding features to perform the class alignment. These observations can also be verified by the qualitative results on the Dark Zurich-val set, as shown in Figure 5. Our ICDA predicts better not only in the large static area such as sky, tree, and sidewalk (red box) but also in small static and dynamic areas such as car, traffic sign, and pole (blue box).

Comparison on BDD100k-night. We also compare our ICDA with other methods on BDD100k-night to verify the generalization. Each model is trained from CityScapes to Dark Zurich-N and directly used to predict the test set of BDD100k-night. As shown in Table 2, our ICDA generalizes best on the BDD100k-night test set and achieves the best per-

Method	mIoU
Deeplab-v2[Chen *et al.*, 2017]	17.3
RefineNet[Lin *et al.*, 2017]	20.4
PSPNet[Zhao *et al.*, 2017]	-
AdaptSegNet[Tsai *et al.*, 2018]	22.0
ADVENT[Vu *et al.*, 2019]	22.6
BDL[Li *et al.*, 2019]	22.8
UDAclustering[Toldo *et al.*, 2021]	20.0
DMAda[Dai and Van Gool, 2018]	28.3
GCMA[Sakaridis *et al.*, 2019]	33.2
MGCDA[Sakaridis *et al.*, 2020]	34.9
CDAda[Xu *et al.*, 2021]	33.8
DANNet[Wu *et al.*, 2021a]	28.0
DANIA[Wu *et al.*, 2021b]	27.0
CCDistill[Gao *et al.*, 2022]	33.0
DAFormer[Hoyer *et al.*, 2022]	34.2
SePiCo[Xie *et al.*, 2022]	36.9
Refign[Bruggemann *et al.*, 2023]	35.2
ICDA(ours)	**38.9**

Table 2: Comparison with the state-of-the-art methods and baseline models on the BDD100k-night test set.

Method	Componets				mIoU
	IE	SE	DAR	PCA	
DAFormer[Hoyer *et al.*, 2022]					31.3
DAFormer with	✓				39.5
DAFormer with	✓	✓			42.0
DAFormer with	✓	✓	✓		43.4
DAFormer with	✓	✓		✓	40.7
Refign[Bruggemann *et al.*, 2023]					43.0
Ours	✓	✓	✓	✓	44.6

Table 3: Ablation studies on Dark Zurich-val set.

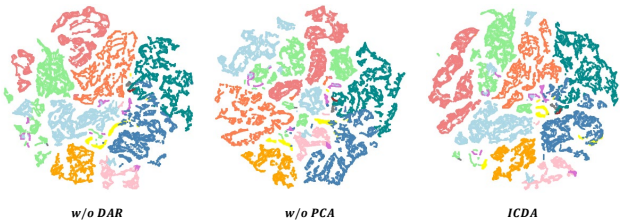

Figure 6: t-SNE analysis of our ICDA and without DAR and PCA.

formance of 38.9%, with an increase of 2%, which shows that our ICDA successfully learns to perceive nighttime scenes by transferring knowledge from the daytime scenes and is robust to nighttime scenes of different conditions.

4.4 Ablation Study

In this section, we perform extensive experiments on several model variants to verify the effectiveness of each proposed component of our ICDA. We measure the performance of each ablated version by evaluating it on the Dark Zurich-val set, and the results are summarized in Table 3.

We first evaluate the backbone DAFormer and get the mIoU of 31.3%. Subsequently, we use illumination enhancement(IE) adopted by the existing methods[Sakaridis *et al.*, 2019; Sakaridis *et al.*, 2020] to generate the stylized nighttime domain and transfer the knowledge from it to the target domain. The performance improves with a gain of 8.2% which shows its effectiveness as proved by the previous methods. Based on this, we then add the spatial enhancement(SE) to form our CEM(no *DAFormer with single SE* since it is meaningless without domain pairs). The performance improves by 2.5%, which verifies SE helps to capture the geometric transformation and improves the robustness. The performance is further improved when we use the DAR to couple the daytime and nighttime images even if without the PCA to transfer the semantic relevance at the class level. This is because our DAR can remedy the predictions of nighttime images thanks to the semantic relevance, which can be seen as the global alignment. However, when performing the class-level alignment between the stylized nighttime domain and target domain without the DAR, the performance is reduced to 40.7%. We hypothesize this is because the target domain features are overfitting to the stylized nighttime domain features without the regularization provided by the global alignment of the DAR. Finally, when all proposed modules are applied, the performance reaches the highest value of 44.6%(higher than 43.0% of Refign[Bruggemann *et al.*, 2023]). By performing the class-level alignment between the coupled domain pairs, our ICDA successfully mitigates the dataset gap and captures the domain-invariant semantic relevance.

4.5 Feature Analysis

To better develop intuition, we draw t-SNE visualizations of the learned feature representations for our ICDA and without the DAR or PCA in Figure 6. With this in mind, we first randomly select an image from the target domain and then map its high-dimensional latent feature representations to a 2D space. From the t-SNE visualizations, we can observe that with our DAR and PCA, (1) the feature representations of the points which belong to the same class are closer to each other. (2) the feature representations are easier to separate in the feature space since the margin between the feature representations of the points which belong to the different classes is larger and fewer feature representations are isolated by others. These two observations further verify the global and class-level alignment function of our DAR and PCA.

5 Conclusions

In this paper, we propose an illumination-coupled domain adaptation framework, which treats the daytime and nighttime with a united view, to address the complex domain gaps of nighttime semantic segmentation. We first use the CEM to construct the source and target domain pairs as the basic adaptation unit. Then we use the DAR to capture the semantic relevance which couples each domain pair at the feature level and adaptively guides the predictions of nighttime images, thus avoiding the illumination gap. To mitigate the dataset gap, we propose the PCA to make use of the category information and perform the fine-grained alignment. Experimental results show our ICDA can largely simplify the complex domain gaps of nighttime semantic segmentation, thus achieving state-of-the-art performance on commonly used benchmark datasets. In future work, we plan to extend the idea of our ICDA to the general adverse environment.

Acknowledgements

This work was supported by the national key R & D program intergovernmental international science and technology innovation cooperation project (2021YFE0101600).

Contribution Statement

Chenghao Dong and Xuejing Kang make equal contributions and share co-first authorship.

References

[Bruggemann *et al.*, 2023] David Bruggemann, Christos Sakaridis, Prune Truong, and Luc Van Gool. Refign: Align and refine for adaptation of semantic segmentation to adverse conditions. In *WACV*, 2023.

[Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[Cordts *et al.*, 2016] 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.

[Dai and Van Gool, 2018] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3819–3824. IEEE, 2018.

[Dai *et al.*, 2017] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[Gao *et al.*, 2022] 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.

[Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hoffman *et al.*, 2018] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018.

[Hoyer *et al.*, 2022] 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.

[Huang *et al.*, 2020] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In *European conference on computer vision*, pages 705–722. Springer, 2020.

[Jiang *et al.*, 2022] Zhengkai Jiang, Yuxi Li, Ceyuan Yang, Peng Gao, Yabiao Wang, Ying Tai, and Chengjie Wang. Prototypical contrast adaptation for domain adaptive semantic segmentation. *arXiv preprint arXiv:2207.06654*, 2022.

[Li *et al.*, 2019] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019.

[Lin *et al.*, 2017] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[Luo *et al.*, 2019] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019.

[Ma *et al.*, 2021] Haoyu Ma, Xiangru Lin, Zifeng Wu, and Yizhou Yu. Coarse-to-fine domain adaptive semantic segmentation with photometric alignment and category-center regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4051–4060, 2021.

[Pan *et al.*, 2020] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773, 2020.

[Rocco *et al.*, 2017] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6148–6157, 2017.

[Sakaridis *et al.*, 2019] 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.

[Sakaridis *et al.*, 2020] 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*, 2020.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[Toldo *et al.*, 2020] Marco Toldo, Andrea Maracani, Umberto Michieli, and Pietro Zanuttigh. Unsupervised domain adaptation in semantic segmentation: a review. *Technologies*, 8(2):35, 2020.

[Toldo *et al.*, 2021] Marco Toldo, Umberto Michieli, and Pietro Zanuttigh. Unsupervised domain adaptation in semantic segmentation via orthogonal and clustered embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1358–1368, 2021.

[Truong *et al.*, 2021] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Warp consistency for unsupervised learning of dense correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10346–10356, 2021.

[Tsai *et al.*, 2018] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018.

[Vu *et al.*, 2019] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[Wang *et al.*, 2020] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *European conference on computer vision*, pages 642–659. Springer, 2020.

[Wang *et al.*, 2021] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8515–8525, 2021.

[Wu *et al.*, 2021a] 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.

[Wu *et al.*, 2021b] 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 & Machine Intelligence*, (01):1–1, 2021.

[Xia *et al.*, 2022] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4794–4803, 2022.

[Xie *et al.*, 2021] 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.

[Xie *et al.*, 2022] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *arXiv preprint arXiv:2204.08808*, 2022.

[Xu *et al.*, 2021] 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.

[Yang and Soatto, 2020] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020.

[Yu *et al.*, 2020] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[Zhang *et al.*, 2021] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[Zhu *et al.*, 2017] 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.