CDAda: A Curriculum Domain Adaptation for Nighttime Semantic Segmentation

Qi Xu, Yinan Ma, Jing Wu, Chengnian Long* and Xiaolin Huang
Department of Automation, Shanghai Jiao Tong University
{txxqsh, yinanma, jingwu, longcn, xiaolinhuang}@sjtu.edu.cn

Abstract

Autonomous driving needs to ensure all-weather safety, especially in unfavorable environments such as night and rain. However, the current daytime-trained semantic segmentation networks face significant performance degradation at night because of the huge domain divergence. In this paper, we propose a novel Curriculum Domain Adaptation method (CDAda) to realize the smooth semantic knowledge transfer from daytime to nighttime. Specifically, it consists of two steps: 1) inter-domain style adaptation: fine-tune the daytime-trained model on the labeled synthetic nighttime images through the proposed frequency-based style transformation method (replace the low-frequency components of daytime images with those of nighttime images); 2) intra-domain gradual self-training: separate the nighttime domain into the easy split nighttime domain and hard split nighttime domain based on the "entropy + illumination" ranking principle, then gradually adapt the model to the two sub-domains through pseudo supervision on easy split data and entropy minimization on hard split data. To the best of our knowledge, we first extend the idea of intra-domain adaptation to self-training and prove different treatments on two parts can reduce the distribution divergence in the nighttime domain itself. In particular, aimed at the adopted unlabeled day-night image pairs, the prediction of the daytime images can guide the segmentation on the nighttime images by ensuring patch-level consistency. Extensive experiments on Nighttime Driving, Dark Zurich, and BDD100K-night dataset highlight the effectiveness of our approach with the more favorable performance 50.9%, 45.0%, and 33.8% Mean IoU against existing state-of-the-art approaches.

1. Introduction

Recent years have witnessed impressive progress in semantic segmentation tasks [15, 18, 6]. However, they are mostly designed to operate in daytime scenes with favorable illumination. The huge domain divergence between daytime and nighttime induces their performance degradation at night. This greatly restricts the application of semantic segmentation algorithms on outdoor scenes which require robust vision systems, such as autonomous driving.

To handle this problem, several domain adaptation works have been proposed to adapt the daytime-trained model to nighttime without labels in the nighttime domain. [10, 26, 29] utilize the twilight domain as the bridge to perform the model adaptation from daytime to nighttime. [31, 24, 26, 29] train the image transferring network to generate the synthetic nighttime images, which can help promote the semantic transfer. The essence of the two lines of work is to introduce the appropriate intermediate domain to realize smooth knowledge transfer. However, all these methods require the additional training data (twilight data) or style transferring network to perform domain adaptation. This is cost-consuming and cannot handle the problem of intra-domain gap.

Therefore we propose a Curriculum Domain Adaptation method (CDAda) to bridge the inter-domain and intra-domain gap together without additional data or network. CDAda separates the domain adaptation process into two steps to realize the progressive from-easy-to-hard domain adaptation, which promotes smoother knowledge transfer. We adopt the Cityscapes dataset [8] and Dark Zurich dataset [29] to realize the domain adaptation. The Dark Zurich dataset contains unlabeled day-night scene image pairs that are coarsely aligned using GPS recordings.

At the step one, inspired from [40], CDAda exploits an improved frequency-based style transformation method. Benefiting from good spatial prior of the used Dark Zurich dataset, daytime images can be translated as synthetic nighttime images by replacing both low-frequency amplitude (reflecting what the target "style" is) and low-frequency phase (reflecting where the target "style" is) of the daytime image with those of the paired night image. Note that the daytime model trained on Cityscapes generates pseudo labels for the Dark Zurich daytime set. Fine-tuning the model with

[*] The corresponding author is Chengnian Long.

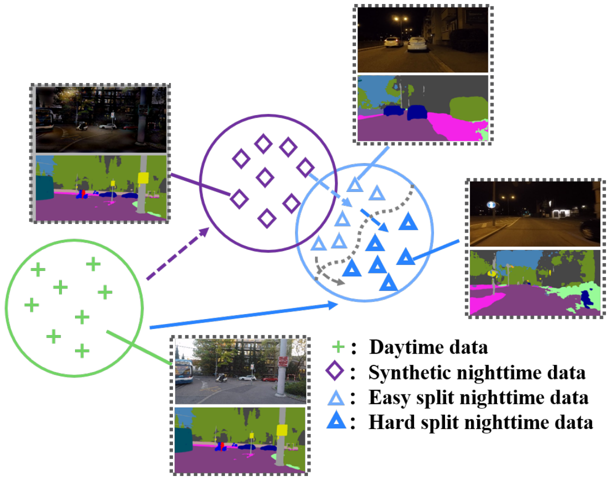

+ : Daytime data
◇ : Synthetic nighttime data
△ : Easy split nighttime data
▲ : Hard split nighttime data

Figure 1: Comparison between direct adaptation (the solid arrow) and CDAda (the dotted arrow). Notation: the dotted gray line represents the threshold separating the nighttime data, which will be dynamically adjusted in the self-training.

the supervision from synthetic nighttime images bridges the inter-domain gap.

At the step two, CDAda proposes an intra-domain gradual self-training which separates the whole process into multiple iteration rounds. In each round, we firstly separate the nighttime data into the easy and hard split based on the “entropy + illumination” ranking principle. Then pseudo labels are generated for the easy split data. The pseudo supervision from these data helps fine-tune the network. As for hard split data with high entropy prediction or too many over-exposed and under-exposed areas, there exists too much noise in the generated pseudo labels of them. Therefore only the exposure-aware entropy minimization is imposed on the pixel-wise prediction. While minimizing entropy, the over-exposed and under-exposed areas are emphasized because these areas often contain seriously degraded visual appearances and structures which affect the model prediction results [32]. Benefiting from different treatments on these two parts of data, semantic knowledge learnt from the easy split data can improve the segmentation performance on the hard split data, which promotes the intra-domain adaptation. In the self-training, we further design a prediction-guidance loss for the used day-night scene image pairs of Dark Zurich dataset. The pseudo labels of the daytime images are adopted as the guidance of segmenting the paired nighttime image. Though pixel-level consistency cannot be realized between the paired images, patch-level consistency can be satisfied well. Therefore, we adopt the pyramid pooling module [43] to process the two paired inputs, then adopt the result of daytime pseudo labels as the supervision of the result of nighttime model prediction to promote the model adaptation.

Generally speaking, CDAda follows the idea of curriculum learning [42]. We adapt the model from easy to hard: daytime → synthetic nighttime → easy split nighttime→ hard split nighttime. The whole adaptation process is shown in Fig. 1. Compared with direct adaptation, this order is more beneficial for model adaptation to realize smoother semantic transfer. In particular, compared with [26, 29], CDAda does not need the additional twilight images as the bridge of inter-domain adaptation and realize the progressive model adaptation from the perspective of reducing the inter-domain and intra-domain gap together .

Our main contributions are summarized as follows:

- We propose a two-step curriculum model adaptation method that follows the appropriate order of model adaptation to realize the smoother semantic knowledge transfer. In particular, we extend the idea of intra-domain adaptation to self-training. It is proved that different treatments on easy-split and hard-split data can promote intra-domain adaptation in the self-training.

- We propose a new style transformation method and prediction-guidance loss based on the good spatial prior of the used day-night image pairs. The prediction of the Dark Zurich daytime image provides the patch-level guidance for segmenting the corresponding Dark Zurich nighttime image. It is shown that these specially designed strategies can significantly enhance the performance of model adaptation.

- Extensive experiments demonstrate CDAda achieves new state-of-the-art segmentation performance on three challenging nighttime scene segmentation datasets, i.e., Nighttime Driving [10], Dark Zurich [26], and BDD100K-night [41] dataset. Ablation study has verified the active effect of each component in CDAda.

2. Related Work

2.1. Nighttime Semantic Understanding

Nighttime is the scene that many vision algorithms must deal with. Therefore, nighttime semantic understanding has attracted numerous attention. Some works realize people detection in nighttime through FIR cameras [39, 12], or visible light cameras [16], or the combination of both [5, 7]. Besides, other works study the detection on the salient objects at night, such as cars [17] and rear lights [30]. Different from the above domain-specific approaches, many works try to design a model which is robust to the changes of illumination [1, 25, 35]. As for the semantic knowledge transfer, [10] shows the twilight images are conductive for the semantic knowledge transfer from daytime to nighttime.

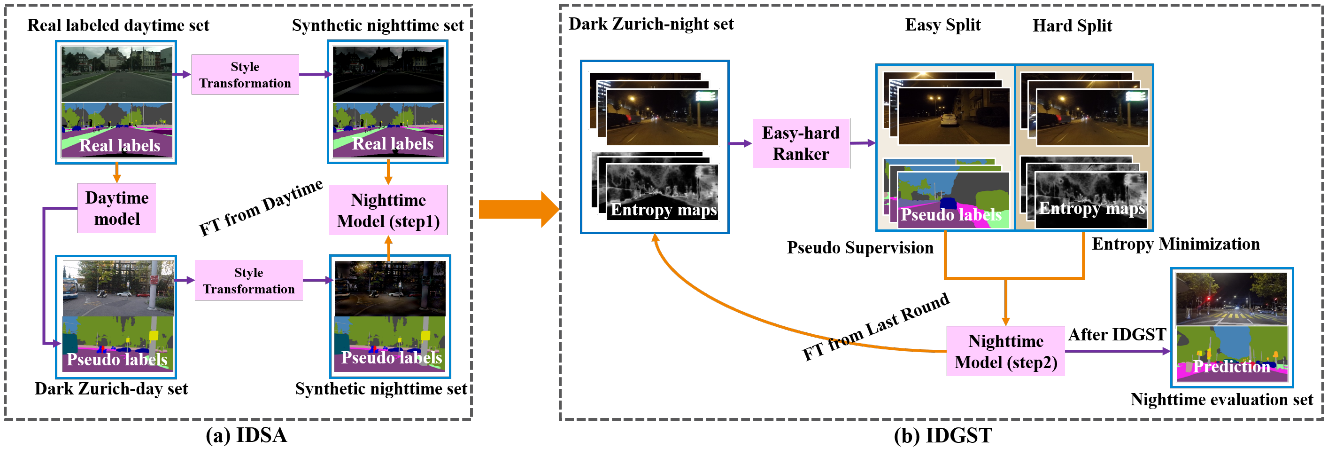


Figure 2: A general overview of the whole framework. (a) is inter-domain style adaptation (IDSA). (b) is intra-domain gradual self-training (IDGST). FT stands for fine-tuning. Orange arrows denote the training process of our model, purple arrows mean the generation of predictions.

[26, 29] extend this idea by learning jointly from unlabeled real nighttime images and synthetic nighttime images. [37] proposes a one-stage domain adaptation network for unsupervised nighttime semantic segmentation through the combination between the image relighting network and the semantic segmentation network. Our work is inspired by the gradual adaptation methods. What differs from the previous methods is that our CDAda needs no additional data or network and realizes the finer division of the adaptation process, which promotes smoother semantic knowledge transfer.

2.2. Model Adaptation

Because the performance of semantic segmentation networks has been rapidly improved, more works [35, 3, 38] turn to study the model adaptation to adverse conditions. [27, 28, 9, 13] learn jointly from the labeled synthetic images and unlabeled real foggy images to adapt the clear-weather model to fog. Nighttime is undoubtedly another important and difficult condition required to be adapted to. Besides, unsupervised domain adaptation based methods were widely proposed to adapt semantic segmentation models from synthetic scenes to real environments [33, 46, 36, 34, 45, 40, 11, 22]. In [46] domain adaptation is defined as an expectation-minimization problem by dynamically iterating between pseudo labeling the unlabeled target data with class-balance curriculum and obtaining a new model with the new generated pseudo labels. IntraDA [22] proposes a self-supervised domain adaptation approach to minimize distribution gap among the target data through adversarial learning. Our work is inspired by the above two method and first extends the idea of intra-domain adaptation to self-training.

2.3. Curriculum Learning

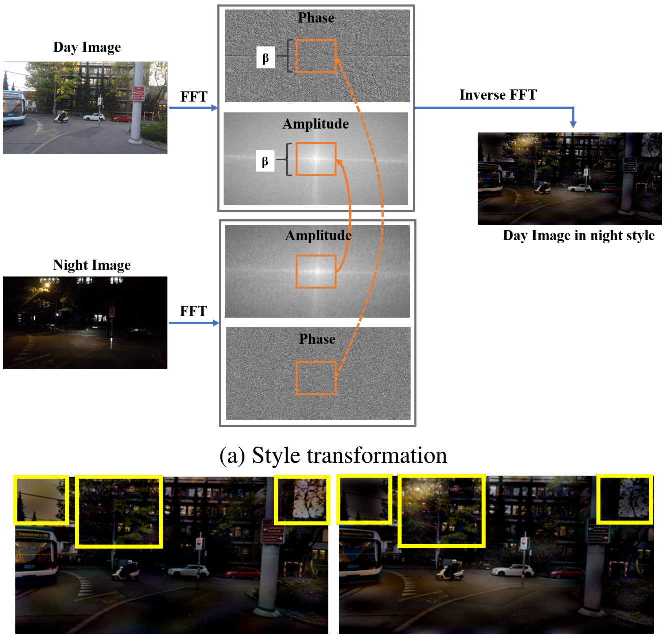
Curriculum learning imitates humans to introduce a learning process, i.e. from easy to complex, which has been proved to promote the optimization of non-convex problems [2]. There exist three types of curriculum in semantic segmentation domain adaptation from the perspective of ordering tasks, utilizing data, and selecting intermediate domains. The first type [42, 20] adds easier tasks than semantic segmentation in curriculum adaptation, which boost the semantic knowledge transfer. The second type [46, 45] interactively generates more pseudo labels for more target data in self-training. The third type [10, 9, 27, 28] gradually transfers semantic knowledge by the transition of the intermediate domain. Our work integrates the idea of the last two types and emphasizes that fine separation of the adaptation process can promote the model adaptation from easy to hard.

3. The Design of CDAda

In this section, we first give an overview of CDAda in Sec. 3.1. Then we explain each step of our CDAda in detail from Sec. 3.2 to Sec. 3.3. Finally we introduce the object function of each step in Sec. 3.4.

3.1. The Overview of Framework

CDAda involves three input datasets: a labeled daytime dataset D_{ld}, an unlabeled daytime dataset D_{ud}, and an unlabeled nighttime dataset D_{un}, which represent Cityscapes, Dark Zurich-day set and Dark Zurich-night set respectively. As shown in Fig. 2, the proposed framework consists of two



(a) Style transformation

(b) Visual results

Figure 3: (a) is the process of style transformation; (b) is the comparison between the synthetic nighttime images generated by FDA [40] (left) and our method (right). Notation: the dotted orange line means the replacement of the low-frequency phase part, which is not performed for style transformation on Cityscapes set.

steps: inter-domain style adaptation (IDSA), intra-domain gradual self-training (IDGST).

3.2. IDSA

Transferring daytime images to the night style is the common method to reduce the huge gap between the daytime domain and nighttime domain. To control the degree of style conversion and reduce the altering of semantic structures, we adopt the frequency-based style transformation method, which is inspired from [40]. Specifically, we first decompose each input image into the amplitude and phase part through the Fast Fourier Transform (FFT), then replace the low-frequency part of the nighttime images into the daytime images before reconstituting the images via the inverse FFT (iFFT) for training. The whole process is illustrated in Fig. 3a.

Different from [40], we conduct a more thorough frequency exchange in CDAda. We find that the replacement of the low-frequency phase part of the daytime image with that of the nighttime image can further facilitate the conversion of style when two images share similar high-level semantics. This is because the low-frequency phase information reflects the position of style information, such as which position should be over-exposed or under-exposed. Luckily the day-night image pairs aligned with GPS recordings are adopted for our training. Therefore the daytime image can learn the position of style information from the paired nighttime image. We propose to replace the amplitude and phase of the daytime image with those of the paired night image, which further boosts the domain adaptation. Note that the phase replacement is only performed in style transformation on the Dark-Zurich day set. Only the amplitude replacement is performed for style transformation on Cityscapes set because of the mismatched spatial prior.

As shown in Fig. 3b, we visualize the synthetic nighttime images produced with only replacing the low-frequency amplitude in FDA [40] and replacing the low-frequency amplitude and phase in our method. It is apparent that the addition of replacing the low-frequency phase makes the synthetic night image better learn the position of exposure in the real night image. The place of the street lamp should be over-exposed, and the place of the sky should be under-exposed, which makes the synthetic images more real.

The frequency-based style transformation converts the style of daytime images x_{ld}, x_{ud} from D_{ld}, D_{ud} to the nighttime, and generates synthetic nighttime images s_{ln}, s_{un} respectively. Note that the parameter β in Fig. 3a means how much low frequency components we replace in daytime images. This reflects the degree of style conversion and we will discuss the effects of different β in Section. 4.4.

3.3. IDGST

Benefiting from good model initialization from IDSA, good pseudo labels of the Dark Zurich-night set can be generated. However, noise in pseudo labels is unavoidable. The noise has two adverse effects: 1) imbalanced class distribution: label generating process is biased to easy-to-adapt categories ignoring other categories; 2) terrible prediction of hard split nighttime data: there exist large regions of some nighttime images on which recognition of the semantic class of the corresponding scene content is hard, even for an experienced human subject [9]. Therefore we take three measures to solve these problems.

Class-balanced Pseudo Labels Generation: inspired from [46], we adopt an offline class-balanced process of generating pseudo labels. Specifically, we rank the confidence of all pixels classified to every class on the data respectively. Then we generate top-confident pseudo labels in every class according to the same ratio α. Because the process of self-training is dynamic, α will be improved respectively in each round. In our work, we follow [46] to set 5 self-training rounds with the ratio $\alpha = 0.2, 0.4, 0.6, 0.8, 1.0$ respectively.

Easy-to-hard Self Training: We propose the gradual adaptation from easy split nighttime images to hard split nighttime images. The hard split data can be selected through two criteria: 1) the entropy of the model prediction; 2) the illumination estimation of the image. Those images with high entropy prediction or too many over-exposure and under-exposure areas can be assumed as hard split. There exists too much noise in these images. Therefore we only impose the exposure-aware entropy minimization on them. Only pseudo labels from easy split data are used for self-training. This effectively reduces the introduction of pseudo label noise from the hard split data at the self-training. In particular, the learned knowledge from easy split data can improve the model prediction on the hard split data, which reflects that different treatments on different parts realize the semantic knowledge transfer between them. This effectively realizes the intra-domain adaptation without adopting adversarial learning [22]. We gradually improve the portion of the easy split part and reduce the portion of the hard split part in each round of self-training. Specifically, the separation score for the unlabeled nighttime image $x_{un} \in \mathcal{R}^{H \times W \times C}$ is defined as

$$Score(x_{un}) = -\frac{1}{H \times W} \sum_{h,w} \sum_{c} p_{un,k}^{h,w,c} log(p_{un,k}^{h,w,c}) + \frac{1}{H \times W} \sum_{h,w} \frac{|I_{un}^{h,w} - 0.5|}{1.0}$$

where $p_{un,k}$ is the corresponding prediction map and I_{un} means the normalized V channel of x_{un} in the hsv space. The former part represents the prediction entropy and the latter part denotes the illumination estimation. In each round, we calculate the separation score for each image x_{un} and rank them according to the corresponding score. Only the top easy split images are selected for generating pseudo labels. To avoid the introduction of additional hyper-parameters, the ratio of selection is also set as α. Such the setting is reasonable. The selection of easy split data is positively related to the selection of confident pixels.

Guidance from the Paired Daytime Image The day-night image pairs are captured from almost identical viewpoints in daytime and nighttime respectively. Therefore a large portion of sharing contents can be utilized to guide the adaptation process through patch-level prediction guidance. The details will be explained in Section 3.4.

3.4. Object Function

In this subsection, we introduce all the objective functions involved in the each step of domain adaptation.

IDSA At the beginning, we generate pseudo labels y_{ud} for D_{ud} through the trained model on D_{ld}. Then we adopt the synthetic nighttime images s_{ln}, s_{un} with the corresponding labels y_{ld} , y_{ud} to fine-tune the model. The training loss is defined as

$$L_{IDSA} = L^c(s_{ln}, y_{ld}) + L^c(s_{un}, y_{ud})$$

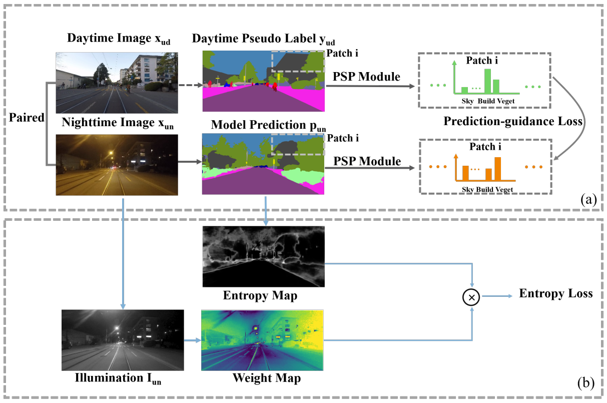


Figure 4: (a) The gray arrow denotes the prediction-guidance loss. (b) The blue arrow means the exposure-aware entropy loss.

where L^c means the cross entropy loss.

IDGST In each round, we split x_{un} as easy split part x_{un}^e and hard split x_{un}^h. Then we generate the class-balanced pseudo labels y_{un}^e for x_{un}^e. The training loss is defined as

$$L_{IDGST} = L^c(s_{ln}, y_{ld}) + L^c(s_{un}, y_{ud}) + L^c(x_{un}^e, y_{un}^e) + \lambda_p L^p(y_{ud}, x_{un}) + \lambda_e L^e(x_{un}^h)$$

where L^p means the patch-level prediction-guidance loss, L^e means exposure-aware entropy loss and λ_p, λ_e is two constants balancing each loss. Next we will explain the two loss in detail:

(a) Prediction-guidance Loss Due to the influence of shooting angle and time, pixel-level alignment cannot be attained but good spatial prior can guarantee preferable patch-level alignment. To avoid controlling the size of the patch, we adopt the pyramid pooling module to get the multi-scale patch-level prediction. As for the output size of the pooling layer, we follow the setting in [43] and set the output size as 1×1, 2×2, 3×3, and 6×6 respectively. We minimize the Kullback–Leibler (KL) divergence between the corresponding output of the one-hot encoding of the daytime pseudo label $y_{ud} \in \mathcal{R}^{H \times W}$ and the model prediction of the paired nighttime image $p_{un} \in \mathcal{R}^{H \times W \times C}$ to realize the patch-level consistency of prediction. As shown in Fig. 4, the prediction-guidance loss is defined as below:

$$L^p(y_{ud}, x_{un}) = \mathbf{KL}(F_{psp}(T_{onehot}(y_{ud}))||F_{psp}(p_{un}))$$

where the F_{psp} means the pyramid pooling module and T_{onehot} means the operation of converting mask to one-hot encoding.

(b) Exposure-aware Entropy Loss The problem that the details of nighttime images in over-exposed and under-exposed areas are destroyed motivates us to make the network pay more attention to the prediction of these areas. As shown in Fig. 4, we impose entropy minimization combining the corresponding weight map on hard split data, which especially minimizes the prediction entropy of the over-exposed and under-exposed areas. The loss is defined as:

$$L^{e}(x_{un})=-(1+\frac{|I_{un}^{h,w}-0.5|}{1.0})\frac{1}{H\times W}\sum_{h,w}\sum_{c}p_{un,k}^{h,w,c}log(p_{un,k}^{h,w,c})$$

4. Experiments

In this section, we first introduce our used model and dataset in Sec. 4.1. Then we explain our training details in Sec. 4.2. After that we compare CDAda with other state-of-the-art model adaptation approaches of nighttime semantic segmentation in Sec. 4.3. Finally we illustrate the effectiveness of CDAda through the ablation experiments in Sec. 4.4.

4.1. Model and Dataset

We adopt the commonly used RefineNet [21] as our choice of architecture for experiments. The publicly available RefineNet-Res101-Cityscapes model which has been trained on the daytime images of Cityscapes dataset is used as the baseline model. The following datasets are used for model training and performance evaluation:

Dark Zurich [9] The Dark Zurich dataset contains 3041 daytime, 2920 twilight, and 2416 nighttime images, which are all unlabeled with the resolution of 1920 × 1080. Especially the images in the three domains are aligned through GPS-based nearest neighbor assignment. These paired images share a large portion of the content, which promotes the domain adaptation between the three domains. In our work, we choose 2416 night-day image pairs to train our curriculum framework **(without using the twilight images)**. Dark Zurich also provides 201 finely annotated nighttime images which are divided into the validation (Dark Zurich-val) and test part (Dark Zurich-test) with 50 images and 151 images respectively. Note that the evaluation of Dark Zurich-test only serves as an online benchmark whose ground truth is not publicly available. In our experiments, we obtained our CDAda on Dark Zurich-test against the annotated ground truths through submitting the prediction results to the online evaluation website.

Nighttime Driving [10] The Nighttime Driving test set provides 50 nighttime images with a resolution of 1920 × 1080. All these 50 images are pixel-wise annotated with the same 19 Cityscapes categories. In our experiments, we also use Nighttime Driving test set for method evaluation.

BDD100K-Night [41] The BDD100K-Night set contains 87 nighttime images with the resolution of 1280 × 720, which are selected from BDD100k and have no obvious errors. The dataset is proposed in [29]. All the images are pixel-wise annotated using the same 19 Cityscapes categories. We adopt BDD100K-Night to further illustrate the model generalization.

4.2. Training Details

We implement our framework in PyTorch [23]. We train the network adopting stochastic gradient descent (SGD) with mini-batch size 1, momentum 0.9, and weight decay 1×10^{-5}. In the whole process we use the constant learning rate of 5×10^{-5}. We train the model for 30,000 iterations for IDSA, 20,000 iterations for every round of the IDGST. Note that though our domain adaptation involves two steps, the training process is similar to the common self-training methods [26, 29]. Therefore our method adds little additional training complexity. In addition, the resized size is set to 1024×512 for the prepossessing of the training data. For the experiments, the hyper-parameters β, λ_p, λ_e are set to 0.01, 1, 0.0025 respectively.

4.3. Comparison with state-of-the-art methods

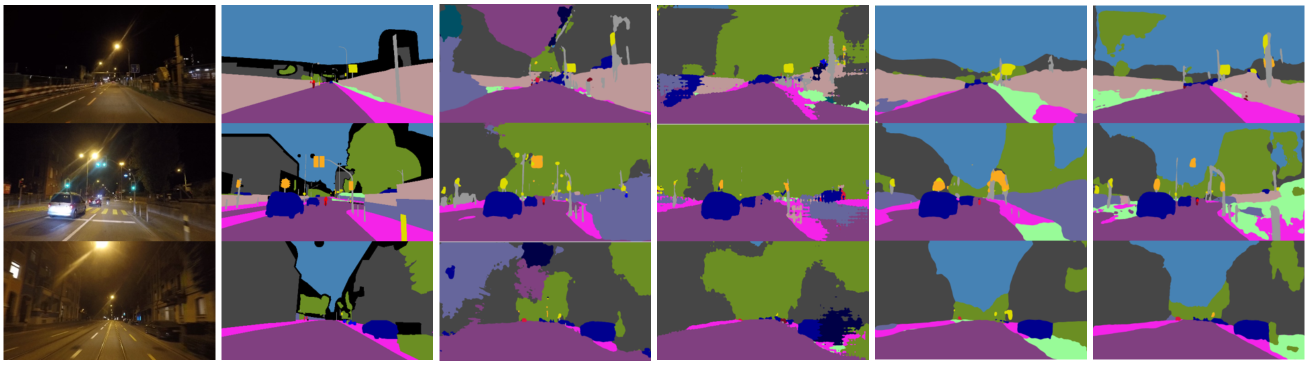
Comparison on Dark Zurich-test Here we will compare CDAda with other state-of-the-art approaches of adapting the semantic segmentation model to nighttime, including AdaptSeg [35], BDL [19], ADVENT [36], DMAda [10], GCMA [26], MGCDA [29] and DANNet [37]. We show the respective mIoU performance in Table 1. The adaptation of AdaptSeg, BDL and ADVENT is trained from Cityscapes to Dark Zurich-night set. Because the used baseline model is different, we also report the performance of the corresponding baseline Cityscapes models for the above methods to conduct a fair comparison. RefineNet is the baseline model of DMAda, GCMA, MGCDA, DANNet and CDAda, while DeepLabV2 [4] is the common baseline model of AdaptSeg, BDL and ADVENT. ResNet-101 [14] is chosen as the backbone of the two baseline models, which allows us to make a comparison directly.

CDAda achieves a 0.7% gain of the overall mIoU over the best score obtained by all existing methods (by DANNet). In particular, we do not require an additional network like DANNet. Compared with MGCDA which adopts the same self-training method, CDAda obtains a significant increase of 2.5% in performance. The respective improvement is apparent in large-scale classes which usually appear dark in the nighttime, such as road, sidewalk. This indicates our method efficiently reduces the domain divergence between daytime and nighttime. Meanwhile, CDAda also gains better performance on some small-scale classes such as pole, bicycle. This indicates that our method can transfer enough semantic knowledge about small-scale classes from daytime to nighttime. To better illustrate the superiority of our method, we show some visual examples in Fig. 5.

Comparison on Nighttime Driving and BDD100K-Night In order to reinforce the generality of our approach, we repeat the above comparison on Nighttime Driving and BDD100K-Night. The respective results are reported in Table 2. Indeed, our method is still by far the best-performing adaptation approach on Nighttime Driving. It is worth men-

Table 1: **Comparison with the state-of-the-art approaches and daytime baseline models on the Dark Zurich-test set.** Notation: the best results are blackened and the second best results are underlined.

Method	**road**	**sidew.**	**build.**	**wall**	**fence**	**pole**	**light**	**sign**	**veget.**	**terrain**	**sky**	**person**	**rider**	**car**	**trunk**	**bus**	**train**	**motorc.**	**bicycle**	mIoU(%)
RefineNet	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
DeepLabV2	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
AdaptSegNet	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	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	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
DMAda	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	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	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(RefineNet)	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	**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**



(a) Image (b) Ground Truth (c) RefineNet (d) AdaptSegNet (e) DANNet (f) CDAda

Figure 5: **Visual results of different approaches on Dark Zurich-val set.**

Table 2: **Comparison with the state-of-the-art approaches and daytime baseline models on the Nighttime Driving set (mIoU1) and the BDD100K-night set (mIoU2).**

Method	mIoU1 (%)	mIoU2 (%)
RefineNet	31.5	26.6
DeepLabV2	32.6	22.9
AdaptSegNet-Cityscapes→ DZ-night	34.5	22.0
ADVENT-Cityscapes→ DZ-night	34.7	22.6
BDL-Cityscapes→ DZ-night	34.7	22.8
DMAda	36.1	28.3
GCMA	45.6	33.2
MGCDA	49.4	**34.9**
DANNet(RefineNet)	42.4	28.2
CDAda	**50.9**	33.8

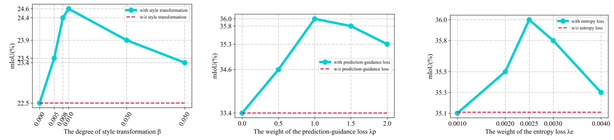
tioning that BDD100K-night dataset is not labeled as elaborately as Dark Zurich-test set and contains some noise though they are selected humanly. In particular these noise mainly reflects in the categories which CDAda performs better than MGCDA, such as sky, vegetation. Even with these issues, our CDAda achieves the second best performance in BDD100K-night (MGCDA achieves the best performance). CDAda improves 19.4% and 7.2% upon RefineNet on Nighttime Driving and BDD100K-night respectively which are a very margin. Though our method has achieved huge improvement on BDD100K, the whole training does not use any data from BDD100K.

Table 3: **Ablation study on several model variants of our method on Dark Zurich-val set.** Notation: CBST means the class balanced self training method in [46]. Exposure Aware means adding the corresponding exposure-aware weight on the entropy loss. Prediction Guidance denotes the addition of prediction-guidance loss.

	Components			mIoU	Gain
initial	Daytime-trained baseline: RefineNet			18.2	
Step One	CycleGAN	FDA	Our FDA	mIoU	Gain
	✓			22.8	+4.6
		✓		23.7	+5.5
			✓	24.6	+6.4
Step Two	The baseline of self-training: CBST			30.6	+12.4
	From Easy to Hard	Exposure Aware	Prediction Guidance	mIoU	Gain
	✓			33.0	+14.8
	✓	✓		33.4	+15.2
	✓	✓	✓	36.0	+17.8

4.4. Ablation Study

Effect of Each Step We measure the active effect of each step and prove the effectiveness of different components through comparing the performance of model variants on the Dark Zurich-val. First, the addition of the synthetic nighttime domain is undoubtedly helpful for the model adaptation. Benefiting from the introduction of no artifacts, FDA [40] brings an additional 0.9% improvement over CycleGAN [44] which is commonly used in other works [31, 24, 26, 29]. Because of favorable spatial prior, we extend the original idea [40] to further swap the low-frequency phase and realize the better style transformation which improves the performance from 23.7% to 24.6%. Second, class balanced self training [46] on the model (step one) improves the performance from 24.6% to 30.6%. Moreover from easy to hard self-training further minimize the intra-domain gap and reduce the introduction of noise from hard split data. This gets the apparent 2.4% improvement in performance. Adopting the exposure-aware entropy loss instead of direct entropy loss (no emphasis on the over-exposed or under-exposed areas) further brings in 0.4% gain. We also see that specially designed prediction-guidance loss further gains 2.6% improvement on the segmentation accuracy through utilizing the spatial similarity of the paired daytime and nighttime images. In general, the full designs of our CDAda bring in an additional 17.8% performance increase than the initial daytime-trained baseline.

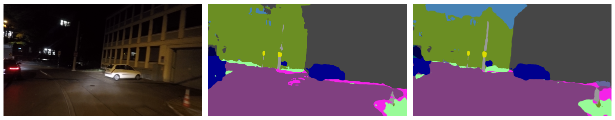


(a) (b) (c)

Figure 6: (a): Performance with the different degree of style conversion β. (b): Performance with the different λ_p. (c): Performance with the different λ_e

Sensitivity to the hyper-parameters We investigate the sensitivity of our method to the hyper-parameters β, λ_p, λ_e and show the results in Fig. 6. From Fig. 6a, it can be seen that in IDSA with the increase of β, the mIoU firstly increases then decreases, illustrating a bell shape curve. The mIoU decreases when β is above a certain threshold, indicating that excessive style shift will harm the adaptation and it is necessary to control the appropriate degree of style shift. With style shift, the optimal mIoU achieved is 2.1% higher than that trained without style shift. As illustrated in Fig. 6b, prediction-guidance loss provides consistent improvement within a wide range of λ_p. From Fig. 6c, we observe that our method is also robust to λ_e within a wide range. Therefore we set $\beta = 0.01$, $\lambda_p = 1$ and $\lambda_e = 0.0025$ for all the experiments.

Intra-domain Adaptation from Easy Split to Hard Split Data Here we discuss that different treatments on the easy split and hard split data can reduce the distribution divergence in the nighttime domain itself. Through pseudo supervision on the easy split data, the model learn more domain-specific knowledge about nighttime, which improves the segmentation performance on the hard split data. As shown in Fig.7, the segmentation performance on the hard split data has been improved a lot from the first round to the second round through comparing the corresponding prediction results. This reflects the semantic knowledge has been transferred from easy split part to hard split part in the nighttime itself. Therefore the intra-domain gap will be gradually reduced in the self-training.



(a) Image (b) Pred (round 1) (c) Pred (round 2)

Figure 7: **Visual comparison between the prediction result of one hard split image at the first round and the second round.**

5. Conclusion

We have introduced a two-step curriculum domain adaptation method for nighttime semantic segmentation. IDSA utilizes labeled synthetic nighttime images to reduce inter-domain gap and provides good initialization for IDGST. IDGST realizes the from-easy-to-hard intra-domain adaptation through class-balanced pseudo supervision on the easy split data and exposure-aware entropy loss on the hard split data. We prove that different treatments on the easy split and hard split data can promote the semantic knowledge transfer between them. In particular, based on the spatial prior of coarsely aligned day-night image pair, the new proposed frequency-based style transformation method and prediction-guidance loss further promote the model adaptation. Compared with other state-of-the-art model adaptation approaches, our CDAda needs no additional training network or training data. Extensive and detailed evaluations with standard IoU on real nighttime sets demonstrate the superiority of our method, which substantially performs favorably against other state-of-the-art methods.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 62136006, 62073215 and 61873166.

References

[1] José M Álvarez Alvarez and Antonio M Ĺopez. Road detection based on illuminant invariance. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):184–193, 2010.

[2] Yoshua Bengio, Jérme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *The International Conference on Machine Learning (ICML)*, 2009.

[3] Mario Bijelic, Tobias Gruber, and Werner Ritter. Benchmarking image sensors under adverse weather conditions for autonomous driving. In *The IEEE Intelligent Vehicles Symposium (IV)*, pages 1773–1779. IEEE, 2018.

[4] 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.

[5] Yuxi Chen and Chongzhao Han. Night-time pedestrian detection by visual-infrared video fusion. In *World Congress on Intelligent Control and Automation*, pages 5079–5084. IEEE, 2008.

[6] Sungha Choi, Joanne T Kim, and Jaegul Choo. Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9373–9383, 2020.

[7] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018.

[8] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[9] Dengxin Dai, Christos Sakaridis, Simon Hecker, and Luc Van Gool. Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. *International Journal of Computer Vision*, 128(5):1182–1204, 2020.

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

[11] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Semi-supervised semantic segmentation via dynamic self-training and class-balanced curriculum. *arXiv preprint arXiv:2004.08514*, 2020.

[12] Junfeng Ge, Yupin Luo, and Gyomei Tei. Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. *IEEE Transactions on Intelligent Transportation Systems*, 10(2):283–298, 2009.

[13] Martin Hahner, Dengxin Dai, Christos Sakaridis, Jan-Nico Zaech, and Luc Van Gool. Semantic understanding of foggy scenes with purely synthetic data. In *The IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 3675–3681. IEEE, 2019.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[15] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 603–612, 2019.

[16] Jong Hyun Kim, Hyung Gil Hong, and Kang Ryoung Park. Convolutional neural network-based human detection in nighttime images using visible light camera sensors. *Sensors*, 17(5):1065, 2017.

[17] Hulin Kuang, Kai-Fu Yang, Long Chen, Yong-Jie Li, Leanne Lai Hang Chan, and Hong Yan. Bayes saliency-based object proposal generator for nighttime traffic images. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):814–825, 2017.

[18] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 9167–9176, 2019.

[19] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6936–6945, 2019.

[20] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 6758–6767, 2019.

[21] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1925–1934, 2017.

[22] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3764–3773, 2020.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[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 *The IEEE Intelligent Vehicles Symposium (IV)*, pages 1312–1318. IEEE, 2019.

[25] German Ros and Jose M Alvarez. Unsupervised image transformation for outdoor semantic labelling. In *The IEEE Intelligent Vehicles Symposium (IV)*, pages 537–542. IEEE, 2015.

[26] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 7374–7383, 2019.

[27] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *The European Conference on Computer Vision (ECCV)*, pages 687–704, 2018.

[28] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018.

[29] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *arXiv preprint arXiv:2005.14553*, 2020.

[30] Ravi Kumar Satzoda and Mohan Manubhai Trivedi. Looking at vehicles in the night: Detection and dynamics of rear lights. *IEEE transactions on Intelligent Transportation Systems*, 20(12):4297–4307, 2016.

[31] Lei Sun, Kaiwei Wang, Kailun Yang, and Kaite Xiang. See clearer at night: towards robust nighttime semantic segmentation through day-night image conversion. In *Artificial Intelligence and Machine Learning in Defense Applications*, volume 11169, page 111690A. International Society for Optics and Photonics, 2019.

[32] Xin Tan, Yiheng Zhang, Ying Cao, Lizhuang Ma, and Rynson WH Lau. Night-time semantic segmentation with a large real dataset. *arXiv preprint arXiv:2003.06883*, 2020.

[33] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7472–7481, 2018.

[34] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1456–1465, 2019.

[35] Abhinav Valada, Johan Vertens, Ankit Dhall, and Wolfram Burgard. Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In *The IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[36] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2517–2526, 2019.

[37] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15769–15778, 2021.

[38] M. Wulfmeier, A. Bewley, and I. Posner. Addressing appearance change in outdoor robotics with adversarial domain adaptation. In *The IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 1551–1558, 2017.

[39] Fengliang Xu, Xia Liu, and Kikuo Fujimura. Pedestrian detection and tracking with night vision. *IEEE Transactions on Intelligent Transportation Systems*, 6(1):63–71, 2005.

[40] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4085–4095, 2020.

[41] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018.

[42] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[43] 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 (CVPR)*, pages 2881–2890, 2017.

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2223–2232, 2017.

[45] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 5982–5991, 2019.

[46] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *The European Conference on Computer Vision (ECCV)*, pages 289–305, 2018.