Cross-Domain Correlation Distillation for Unsupervised Domain Adaptation in Nighttime Semantic Segmentation

Huan Gao[1] Jichang Guo[1]* Guoli Wang[2] Qian Zhang[2]
[1]School of Electrical and Information Engineering, Tianjin University.
[2]Horizon Robotics.
{gh99, jcguo}@tju.edu.cn, {guoli.wang, qian01.zhang}@horizon.ai

Abstract

The performance of nighttime semantic segmentation is restricted by the poor illumination and a lack of pixel-wise annotation, which severely limit its application in autonomous driving. Existing works, e.g.*, using the twilight as the intermediate target domain to perform the adaptation from daytime to nighttime, may fail to cope with the inherent difference between datasets caused by the camera equipment and the urban style. Faced with these two types of domain shifts,* i.e.*, the illumination and the inherent difference of the datasets, we propose a novel domain adaptation framework via cross-domain correlation distillation, called CCDistill. The invariance of illumination or inherent difference between two images is fully explored so as to make up for the lack of labels for nighttime images. Specifically, we extract the content and style knowledge contained in features, calculate the degree of inherent or illumination difference between two images. The domain adaptation is achieved using the invariance of the same kind of difference. Extensive experiments on Dark Zurich and ACDC demonstrate that CCDistill achieves the state-of-the-art performance for nighttime semantic segmentation. Notably, our method is a one-stage domain adaptation network which can avoid affecting the inference time. Our implementation is available at* https://github.com/ghuan99/CCDistill.

1. Introduction

Semantic segmentation as one of the fundamental topics in computer vision, has been widely used in many critical downstream tasks [4, 12]. While a large variety of approaches have been proposed [2, 30], they are predominantly designed to train on daytime images with favorable illumination. However, outdoor applications require satisfactory performance in more challenging scenes, such as nighttime. In this work, we focus on semantic segmentation at nighttime, which is primarily limited by the low exposure of the captured images and the lack of ground truth.

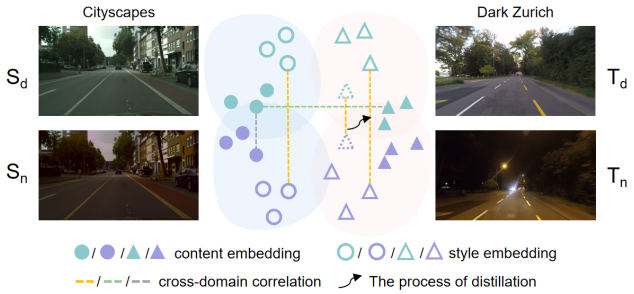

Figure 1. Every two embeddings connected by the dotted line come from two domains, and they only have one difference in illuminance (*i.e.*, each column) or dataset (*i.e.*, each row). The cross-domain correlation reflects the similarity of the two domains, and can also be considered as a concrete representation of the domain shift. Here we only illustrate the cross-domain style distillation. Our main idea is to make the different cross-domain correlations under the same domain shift consistent.

To handle this problem, many domain adaptation methods have been proposed to adapt the daytime-trained model to nighttime without requiring ground-truth labels in the nighttime domain. In [34–36, 38, 48], they apply an image transfer network to stylize daytime or nighttime images and generate synthetic datasets. However, the style transfer network cannot fully utilize the semantic embedding of the segmentation task and also increases the inference time. Some works [7, 35, 36] utilize the twilight as the intermediate target domain. These methods require additional training data and the training process is complex. Most importantly, all these methods ignore inherent difference between datasets, treating daytime images from different datasets as the same style. Prior work [13] points out that appearance discrepancy has a significant impact on the effect of adaptation. Ignoring the inherent difference can adversely affect

*Corresponding author

domain adaptation.

Considering the illumination and inherent difference between labeled daytime images and unlabeled nighttime images, we intend to construct an end-to-end multi-source multi-target domain adaptation framework for nighttime semantic segmentation (shown in Fig. 1). The Dark Zurich [35] containing unlabeled daytime (T_d) and nighttime (T_n) image pairs and Cityscapes [6] containing labeled daytime images (S_d) are adopted as our datasets. It can be seen from Fig. 1, that T_d and T_n are taken at different times in the close scene, thus there is the huge difference of illumination but highly overlapped semantic information. Although S_d and T_d are both daytime images, there are obvious differences in the urban style and color tone. We treat the difference in illumination and dataset as the domain shift.

There is a wide literature on knowledge distillation works [9, 20, 24, 39, 44, 51] that have explored the cross-modal learning. One of strategies in these methods is to exploit the semantic consistency of images across domains as prior knowledge [39, 44]. However, most of them [9, 22, 39] focus on one teacher and one student. As illustrated in Fig. 1, we observe that if we can get the S_n with content of S_d and illumination style of T_n, the degree of difference in content between S_d and T_d should be consistent with that between S_n and T_n. Similarly, the degree of difference in illumination or content between S_d and S_n is consistent with that between T_d and T_n. Therefore, we can leverage the invariance of domain shifts as prior knowledge to implement knowledge distillation in multi-source multi-target domain.

With this insight, we propose a cross-domain correlation distillation approach, which is implemented on the content and style knowledge contained in the feature. The degree of cross-domain difference is obtained by the similarity of two content or style embeddings with only one domain shift, and it can also be regarded as a concrete representation of the domain shift. The cross-domain content correlation is utilized to realize the knowledge distillation from the labeled daytime to the unlabeled nighttime domain, so as to improve the performance of the nighttime semantic segmentation. The premise for the effectiveness of the cross-domain content distillation is that the generated and real nighttime images tend to be as consistent as possible in style. Therefore, we first employ a simple image translation method [13] to align holistic distribution on LAB color space to initially reduce the style discrepancy between day and night. And the cross-domain style distillation can further achieve the style transfer at the semantic-level.

Different from reducing the illumination shift adopted by previous works, it is possible to obtain accurate features of nighttime images by exploiting the consistency of domain shift. We evaluate the performance of CCDistill on Dark Zurich [35], ACDC [37] datasets. Our main contributions are summarized as follows:

- For nighttime semantic segmentation, we propose an end-to-end unsupervised domain adaptation framework, CCDistill, which requires neither extra data nor style transfer network, thus it does not affect the inference time of the semantic segmentation network.
- We propose the cross-domain correlation distillation algorithm, which utilizes the invariance of domain shifts to perform knowledge distillation on content and style embeddings separately. It enables knowledge distillation to be free from the adverse effect caused by the complex domain shifts.
- Extensive experiments on the Dark Zurich and ACDC datasets verify that our network achieves a new state-of-the-art performance of nighttime semantic segmentation.

2. Related works

Domain adaptation. Domain adaptation can effectively tackle the inconsistent data distribution in different domains. A line of methods utilize the principle of model consistency to reduce the data distribution gap by data augmentation [32]. Chen *et al*. [3] combine source and target domain by the cutmix [54] and concat. And [33] holds the view that the input level does not follow the cluster assumption, which can be maintained in the embedding space. Therefore, they add different perturbations to the output of the encoder.

However, domain adaptation methods based on style transfer are often more intuitive and integrated [16, 49]. In [13, 31], they both convert the source domain image to the LAB color space for style translation. Isobe *et al*. [21] convert all other domains into the style of the current target domain for further training.

Instead of using data augmentation or style transfer, designing the loss function to constrain the data distribution can also achieve feature alignment [10, 18, 19, 23, 26, 53]. Wang *et al*. [47] apply the projection head to map the feature to a 256-d l2-normalized vector, and use the NCE loss on the mapped vector to explore the global semantic relationship. Liu *et al*. [29] utilize the KL divergence on the mean and variance stored in the BN layer of the model to make the data distributions similar to each other.

Taking into account the characteristics of the nighttime semantic segmentation, the general domain adaptation methods may fail to cope with the complex domain shift between the daytime and nighttime domains. Therefore, we combine the latter two strategies to construct multi-source and multi-target domains through image-level and semantic-level style transfer, and obtain content embedding by using the JS divergence to constrain data distribution.

Knowledge distillation . In knowledge distillation (KD), the goal is to transfer additional feedback from the teacher

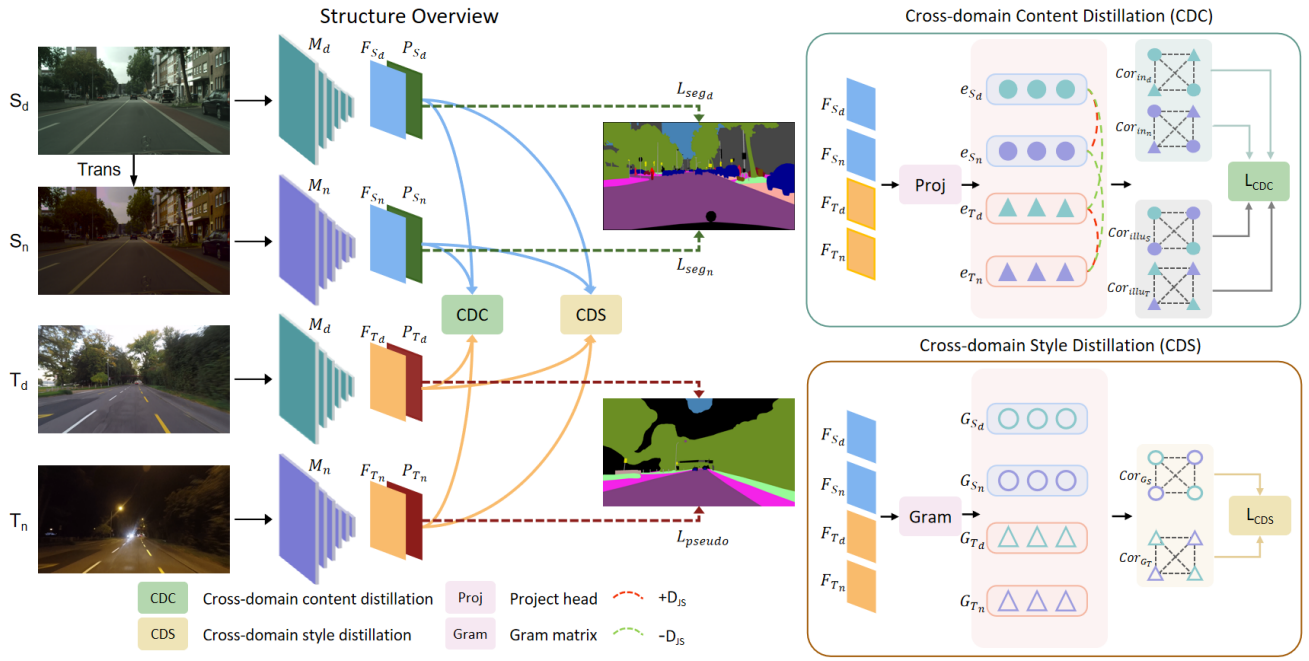

Figure 2. Framework. 1) The overview of our proposed framework is shown on the left. The architecture consists of two semantic segmentation models M_d and M_n. The colored solid arrows represent the data flow of the middle layer features $\mathbf{F}_D$ from different domains, and the colored dash arrows represent the supervision to the outputs $\mathbf{P}_D$. 2) The specific distillation is shown on the right. The $\mathbf{e}_D$ and $\mathbf{G}_D$ represent content and style embedding, respectively. For $\mathbf{F}_D$, $\mathbf{P}_D$, $\mathbf{e}_D$ and $\mathbf{G}_D$, the subscript indicates which domain they are obtained from.

to the student. In early KD methods [15], the knowledge transfer is implemented by minimizing the *Kullback-Leibler* (KL) divergence between the predicted distribution of the student and teacher. Recent studies have explored cross-model KD, which transfers high-level knowledge across different modalities [8, 17, 20, 24, 39, 44, 45, 50].

The IntRA-KD [17] calculates the mean, variance, and skewness of each category in the feature as the statistics of the current distribution, and uses the cosine similarity of the moment vectors to perform distillation. Similarly, [8] applies the Euclidean distance to represent the correlation between instances. And [28, 44] realize pair-wise distillation by dividing the feature into several nodes and then calculating the similarity between different nodes.

Inspired by the above methods, we explore to make use of the correlation contained in features. The aforementioned methods mostly focus on the situation of a single teacher and a single student. Instead, there are multiple domains in our task, and the inputs of different models include differences in illumination and datasets. Hence, we adopt the distance of embeddings from the two domains with only one kind of domain shift as the high-level representation to transfer knowledge.

Nighttime semantic segmentation. Previous works on nighttime semantic segmentation apply adversarial models to achieve the style translation from daytime to nighttime [34–36, 38, 48]. In order to deal with the domain gap, DANNet [48] uses a style translation network to transform different domains as the same style. Besides RGB images, HeatNet [42] additionally uses thermal data that is not sensitive to illumination. Many methods adopt twilight as the intermediate domain to gradually reduce the distribution discrepancy [7, 35, 36]. And these methods either require extra data, or need to design additional networks that affect the inference time, and the training process is complicated. Therefore, instead of reducing the illumination shift between labeled daytime and unlabeled nighttime images with a style transfer network, we leverage the domain shift and regard the cross-domain correlation as the concrete representation of the domain shift to realize domain adaptation.

3. Method

3.1. Problem Formulation

Most existing nighttime semantic segmentation methods mainly consider illumination difference and achieve domain adaptation by reducing the illumination shift between S_d and T_n. While based on our observation, the domain shifts between S_d and T_n include not only illumination differ-

ence but also inherent difference between datasets caused by camera equipment and urban appearance. Regardless of which domain shift is ignored, the effect of domain adaptation will be adversely affected.

In this section, through constructing the multi-source multi-target domain adaptation network, we can select two domains with only one domain shift and calculate the degree of difference between the two domains. Then we propose the cross-domain correlation distillation by using the invariance of cross-domain difference to achieve domain adaptation. Formally, our network involves a source domain S_d, a synthetic dataset served as another source domain S_n, and two target domain T, denoted as T_d and T_n, where $D \in \{S_d, S_n, T_d, T_n\}$ and these four elements represent Cityscapes (daytime), Cityscapes (synthetic nighttime), Dark Zurich (daytime) and Dark Zurich (nighttime), respectively. Note that only images from S_d and S_n have the pixel-wise annotation. And T_d and T_n are taken at different times in the same scene.

The overall architecture of our proposed method is shown in Fig. 2. Our algorithm has four major components:

- Semantic Segmentation network. We adopt the RefineNet [27] as the semantic segmentation network, training two segmentation models M_d and M_n simultaneously, where M_d takes S_d and T_d as inputs, and M_n takes S_n and T_n as inputs. Our goal is to get the accurate prediction map $\mathbf{P}_{T_n}$ for T_n without using the pixel-level annotation.
- Project Head. This block is implemented as two 1x1 convolutional layers with ReLU [47]. The intermediate feature $\mathbf{F}_D$ of M_d or M_n is input to the project head, and it is mapped to the 256-d l2-normalized vector to extract content embedding for knowledge distillation. Note that this block is only utilized during training, thus the inference time will not be affected.
- Cross-domain content distillation. We adopt cosine similarity between two content embeddings $\mathbf{e}_D$ from different domains to represent the degree of content difference.
- Cross-domain style distillation. Different from the content knowledge, style embedding $\mathbf{G}_D$ is obtained by calculating the Gram matrix [11] of the feature $\mathbf{F}_D$ itself, and we also use similarity function to measure the cross-domain style difference.

3.2. Cross-domain correlation distillation

It can be seen from Fig. 2, in addition to the discrepancy of illumination between daytime and nighttime, the daytime images from different datasets also have its particular color tone and urban style. If we can get the S_n with content of S_d and illumination style of T_n, the degree of difference in content between S_d and T_d should be consistent with that between S_n and T_n. Similarly, the degree of difference in illumination or content between S_d and S_n is consistent with that between T_d and T_n. This invariance of difference in illumination or content can be exploited as prior knowledge to guide the model to extract accurate features for T_n.

Motivated by the cross-model knowledge distillation [8, 39, 44], we propose the cross-domain content distillation (CDC) and cross-domain style distillation (CDS). The former conducts the transfer of content knowledge which is essential for the segmentation task, and the latter realizes the style transfer in semantic level.

The following subsections describe in detail how content and style embeddings are extracted and how the degree of difference between the two domains is calculated.

Cross-domain content distillation. The same image always maintains the same semantic content in different styles. Similarly, two images from different datasets should maintain the degree of content difference across styles. The CDC exploits this invariance of content difference to perform semantic knowledge distillation.

Due to the difference in the input of the model M_d and M_n, the feature distribution differs from each other. Here we first utilize the project head to map the features $\mathbf{F}_D$ into the common embedding space, and get $\mathbf{e}_D$. Then we further introduce the *Jensen-Shannon* (JS) divergence to constrain the feature distribution. Specifically,

$$L_{JS} = \lambda(JS(\mathbf{e}_{S_d}||\mathbf{e}_{S_n}) + JS(\mathbf{e}_{T_d}||\mathbf{e}_{T_n}))- (JS(\mathbf{e}_{S_d}||\mathbf{e}_{T_d}) + JS(\mathbf{e}_{S_n}||\mathbf{e}_{T_n})) \quad (1)$$

In order to get the content knowledge contained in the feature, the distribution of embeddings with the same semantic information needs to be close, as the first term in Eq.(1). And at the same time, it is necessary to ensure that the embeddings with different semantic information keep a certain distance, as the second term in Eq. (1). λ is the coefficient used to control the effect of reverse JS divergence and it is set to 4.

After getting the content embedding $\mathbf{e}_D$, we adopt the similarity function to express the degree of content difference between the two domains. In this way, we can get the cross-domain content knowledge, which is formulated as:

$$Cor_{illu_k} = cos(\mathbf{e}_{k_d}, \mathbf{e}_{k_n}), k \in \{S, T\} \quad (2)$$

$$Cor_{in_r} = cos(\mathbf{e}_{S_r}, \mathbf{e}_{T_r}), r \in \{d, n\} \quad (3)$$

Cor_{illu_k} indicates the correlation of the content within the domain S or the domain T, and Cor_{in_r} indicates the inherent correlation between different datasets in the daytime or nighttime scene. $cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T\mathbf{y}}{||\mathbf{x}||_2||\mathbf{y}||_2}$ is the commonly used cosine similarity. Intuitively, the model which takes the daytime images as input tends to be less difficult to train, and the ground truth in the domain S can also be helpful to extract more superior features. Therefore, Cor_{illu_S} is used to guide Cor_{illu_T}, and Cor_{in_d} is used to guide Cor_{in_n}. We

utilize the cross-domain correlation to realize the knowledge transfer from domain S to domain T, from daytime to nighttime, and reduce the disparity of model performance. Cor_{illu} and Cor_{in} represent the patch-level correlation, so they are still effective even if there is the parallax between T_d and T_n. The cross-domain content distillation loss is given as follows:

$$L_{CDC} = ||Cor_{illu_S} - Cor_{illu_T}||_2^2 + ||Cor_{in_d} - Cor_{in_n}||_2^2 + L_{JS} \quad (4)$$

The domain shifts that exist between these four domains can be divided into two categories: illumination and inherent difference between different datasets. We select two domains with only one kind of shift each time, construct their correlation graph. For example, Cor_{in_d} reflects the similarity between the content of S_d and T_d, and there is only the inherent difference between S_d and T_d. Similarly, Cor_{in_n} reflects the similarity between S_n and T_n, and they also have only the inherent difference. Cor_{in_d} and Cor_{in_n} can be regarded as concrete representations of the inherent difference in the content of daytime and nighttime images between datasets, respectively. Therefore, the process of forcing Cor_{in_d} and Cor_{in_n} to be equal, as the second term in Eq. (4), is to utilize the invariance of inherent difference between datasets to achieve the knowledge distillation while avoiding the adverse effects caused by the other kind of domain shift. In a similar way, Cor_{illu} takes advantage of the invariance of content in the same dataset.

Cross-domain style distillation. The premise to implement the CDC is to be able to generate S_n with the same illumination style as T_n. Previous approaches [35, 36] generate nighttime images through style translation models, *e.g.*, CycleGAN [55], yet the semantic features in segmentation task are underutilized.

We first align the mean and variance of S_d with T_n in the LAB space to get S_n [13, 31]. This pre-process can realize the holistic style transformation and decrease the difficulty of model convergence. However, for nighttime images, due to the presence of traffic lights, headlights, etc., there is local overexposure of brightness. If only this holistic style transformation is performed, the generated image will still be quite different from the real nighttime image. As shown in Fig. 2, after we perform the moment match in the LAB space, the tone of S_n has tended to T_n at the holistic level. But for the underexposed or overexposed areas in T_n, the effect of this style transformation is still not satisfactory. Therefore, we propose the cross-domain style distillation (CDS) to further achieve semantic-level style transfer during the training of segmentation model.

In style transfer [5, 11, 46], the Gram matrix is used to indicate the self-correlation of features in the channel dimension, which consists of the correlation between the responses of different filters. We adopt the Gram matrix $\mathbf{G}_D \in R^{C \times C}$ to represent the style of the feature, where $\mathbf{G}_D$ is the inner product of the vectorised feature maps of $\mathbf{F}_D$ on channel i and j respectively:

$$\mathbf{G}_D = \sum_p \mathbf{F}_D^{ip} \mathbf{F}_D^{jp}, D \in \{S_d, S_n, T_d, T_n\} \quad (5)$$

where p is the pixel of $\mathbf{F}_D$. After obtaining the style knowledge of the feature $\mathbf{F}_D$ itself, the principle of our style transfer is similar to that of the CDC. We also build the cross-domain style graph, and this can be formulated as:

$$Cor_{G_k} = cos(\mathbf{G}_{k_d}, \mathbf{G}_{k_n}), k \in \{S, T\} \quad (6)$$

$$L_{CDS} = ||Cor_{G_S} - Cor_{G_T}||_2^2 \quad (7)$$

Cor_{G_k} reflects the degree of illumination difference in the source or target domain. The alignment of style difference between domain S and T achieves the semantic-level style transfer, as defined in Eq. (7).

It is worth noting that only the style correlation in the domain S or the domain T is used here. The main reason is that although there is inherent style shift between S_d and T_d, this is less noticeable than the illumination difference between the daytime and nighttime, thus the correlation of the Gram matrices between them is not strong. The CDS mainly utilizes the invariance of illumination difference between daytime and nighttime images in the same dataset to perform style transfer, so that the S_n gradually approaches the illumination style of T_n. And this choice can also exclude the adverse effect caused by the inherent difference.

3.3. Objective functions

In summary, the total loss of our method is written as follows:

$$L = L_{seg_n} + L_{seg_d} + L_{pseudo} + \lambda_1 L_{CDC} + \lambda_2 L_{CDS} \quad (8)$$

where L_{seg_n} is the weighted cross-entropy loss between the prediction map $\mathbf{P}_{S_n}$ and the corresponding ground truth, and L_{seg_d} is in the same way. L_{pseudo} is the static loss [48], which uses the predictions of static object categories for the daytime images T_d as the pseudo labels to provide pixel-level supervision on T_n. The λ_1, λ_2 are hyper-parameters that balance the influence of distillation losses on the main task, which are set to 2 and 1, respectively.

4. Experiments

4.1. Datasets

The following datasets are used for model training and performance evaluation:

Cityscapes [6] is an autonomous driving dataset captured from street scenes in 50 cities, with pixel-wise annotations of 19 semantic categories. It contains 2,975 images for training, 500 images for validation and 1,525 images for testing. All images are at a fixed resolution of 2,048×1,024.

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

Method	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	mIoU
DeepLab-v2-Cityscapes [1]	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-Cityscapes [27]	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
AdaptSegNet-Cityscapes→DZ-night [41]	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-Cityscapes→DZ-night [43]	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-Cityscapes→DZ-night [25]	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-Cityscapes→DZ-night [40]	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 [7]	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 [35]	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 [36]	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) [48]	**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
Ours	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**

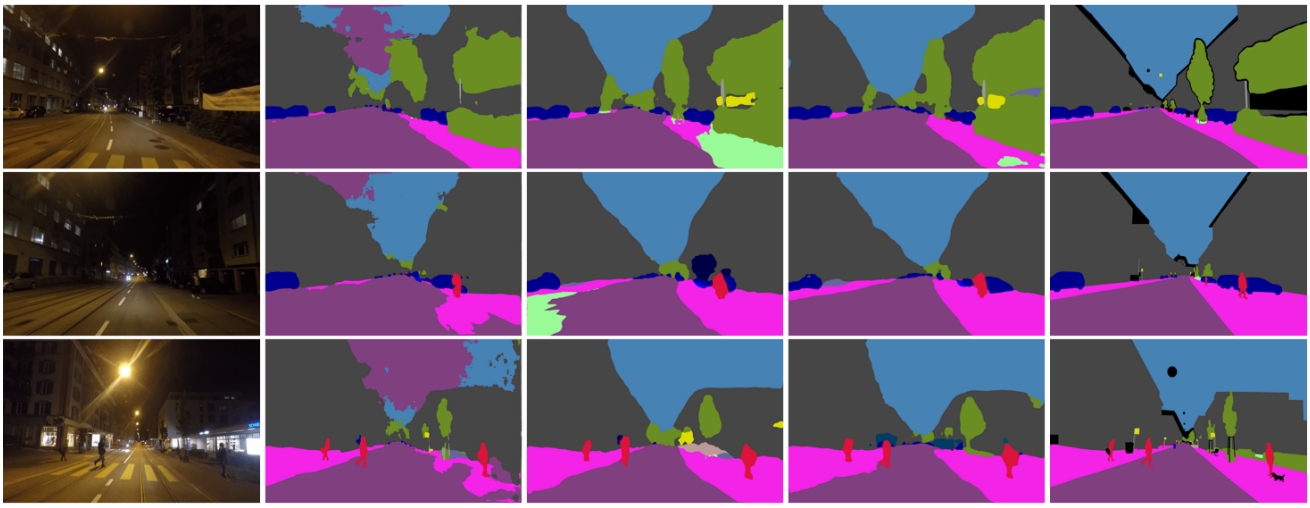

Figure 3. The qualitative comparison between our approach and some existing state-of-the-art methods on the Dark Zurich-val set.

In this paper, we adopt the Cityscapes training set in the training of our network.

Dark Zurich [35] is captured in Zurich, with 3,041 daytime, 2,920 twilight and 2,416 nighttime images for training, which are all unlabeled with a resolution of 1,920x1,080. Each nighttime image has a corresponding daytime image as auxiliary, which constitutes a data pair that can be used for the knowledge distillation in our proposed network. Thus we use the 2,416 night-day image pairs in our training process. The Dark Zurich also contains 201 manually annotated nighttime images, of which 151 (Dark Zurich-test) are used for testing and 50 (Dark Zurich-val) are used for validation. Note that the evaluation of Dark Zurich-test only serves as an online benchmark, and its ground truth is not publicly available.

ACDC [37] consists of 4006 images including four common adverse conditions: fog, rain, nighttime and snow. The images under nighttime scenes have pixel-wise annotations, and are further divided into 400 training, 106 validation and 500 test images. This dataset and the Dark Zurich are both proposed by Sakaridis *et al*. [35], thus it shares the similar style and appearance with the Dark Zurich. So we adopt the ACDC-night-val to further evaluate the effect of our network for domain adaptation.

4.2. Implementation details

We implement the proposed network using PyTorch on a single Titan RTX GPU. We adopt the RefineNet [27] as our semantic segmentation model, which is pre-trained on the Cityscapes dataset with the ResNet-101 [14] as backbone. Both our models are trained by the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}, and the initial learning rate is set as 2.5×10^{-4}. Then the learning rate is decreased with

Table 2. Comparison with the state-of-the-art methods and baseline models on the ACDC-night-val set (mIoU1) and the BDD100K-night set (mIoU2).

Method	mIoU1	mIoU2
DeepLab-v2-Cityscapes [1]	16.3	17.3
RefineNet-Cityscapes [27]	20.3	20.4
AdaptSegNet-Cityscapes→DZ-night [41]	23.8	22.0
ADVENT-Cityscapes→DZ-night [43]	26.2	22.6
BDL-Cityscapes→DZ-night [25]	23.9	22.8
UDAclustering-Cityscapes→DZ-night [40]	24.5	20.0
DMAda [7]	-	28.3
GCMA [35]	-	33.2
MGCDA [36]	29.0	**34.9**
DANNet(RefineNet) [48]	37.0	30.3
Ours	**37.7**	33.0

Table 3. Ablation studies of our proposed method on Dark Zurich-test set.

Method	mIoU
RefineNet	28.5
w/o CDC	45.3
w/o project head in CDC	38.2
w/o L_{JS} in CDC	44.6
w/o illuminance correlation in CDC	45.7
w/o inherent correlation in CDC	45.9
w/o CDS	44.0
w/o LAB-based Trans	43.5
w/o CDC and CDS	44.8
Ours	47.5

the poly policy with a power of 0.9. The batch size is set to 2. The total number of training iterations is 50k. Following [48], we apply random cropping with a crop size of 512 for Cityscapes dataset, and with a crop size of 960 for Dark Zurich which is then resized to 512. At the inference time, there is no any change introduced to the final model M_n.

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

Comparison on Dark Zurich. We compare our proposed method with some existing state-of-the-art methods, including DMAda [7], GCMA [35], MGCDA [36], DANNet [48], and several other domain adaptation approaches [25, 40, 41, 43] on Dark Zurich-test. The MGCDA, GCMA, DMAda and DANNet adopt the RefineNet [27] as the baseline, while other methods use the Deeplab-v2 [1]. To ensure a fair comparison, we perform our method on the RefineNet. Note that both the baselines use the ResNet-101 [14] as backbone. Table 1 shows the quantitative comparison with other methods on Dark Zurich-test. The mIoU is calculated by the average of the intersection-over-union (IoU) among all 19 categories.

Our method surpasses the existing methods with around 3.2% increase on mIoU. In particular, CCDistill is a one-stage adaptation framework with requiring no additional network in the inference. We also observe that our approach has comparable effects in all large-scale categories such as terrain, sidewalk and road, which proves that our method achieves the style transfer from daytime to nighttime and thus realizes the cross-domain knowledge distillation. Moreover, CCDistill significantly improves the performance of categories with relatively few occurrences, such as train and motorcycle. This also indicates that our method does transfer the semantic-level correlation knowledge. The qualitative results on Dark Zurich-val, as shown in Fig. 3, can also verify this observation.

Comparison on ACDC. In order to verify the effectiveness of the proposed model on nighttime semantic segmentation, we further conduct comparative experiments on the ACDC-night-val, and the results are shown in Table 2. The ACDC-night has a similar nighttime style with the Dark Zurich-night, so it is reasonable that the CCDistill achieves the best performance on ACDC-night-val, and a 0.7% improvement of mIoU is gained. The visualization comparison on ACDC-night-val is shown in Fig. 4.

4.4. Generalization test

Same as the daytime images, there are also domain shifts between nighttime images from different datasets. In order to verify the generalization of our proposed method, we also compare with other methods on the BDD100K-night. The BDD100K-night contains 87 images with the resolution of 1,280×720, which is manually selected by [36] from the 345 nighttime images of BDD100K [52]. The appearance and lighting tone between Dark Zurich and BDD100K-night are quite different. As shown in Table 2, even though the target domain of our proposed method is Dark Zurich and it is the domain shift between Cityscapes and Dark Zurich that we utilize, we still get the comparable performance on the BDD100K-night.

4.5. Ablation study

In this section, extensive experiments on several model variants are conducted to verify the effectiveness of each proposed component. We measure the performance of each ablated version by evaluating it on the Dark Zurich-test. Results are summarized in Table 3.

The content correlation across domains is the core of the knowledge distillation in our method. We set up five ablated version to prove the effect of the proposed CDC. First, re-

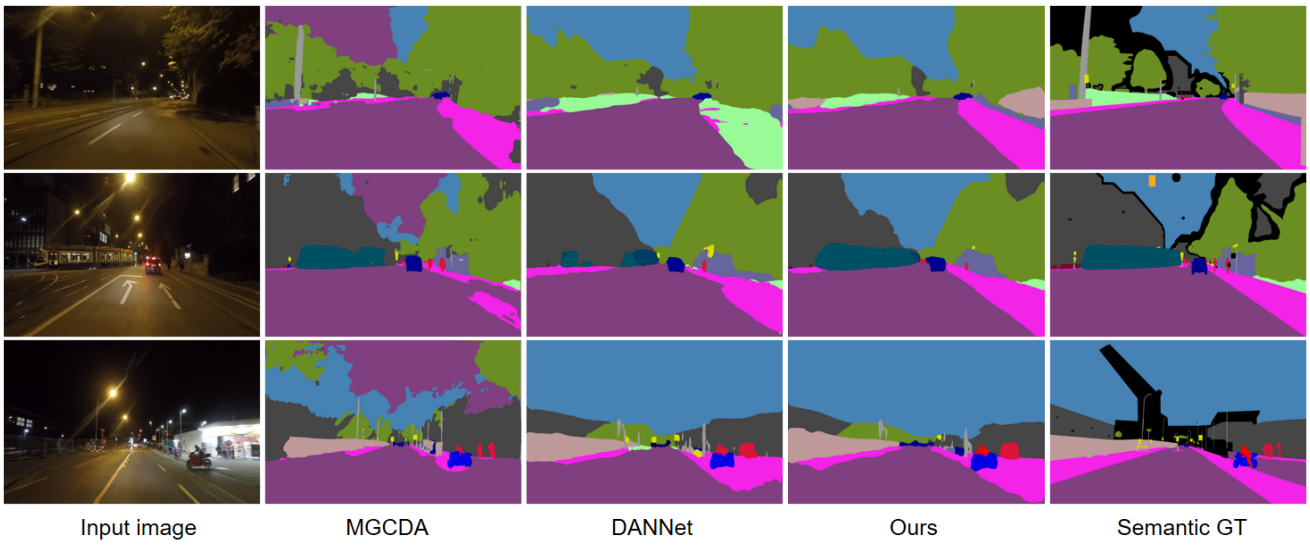

Figure 4. The qualitative comparison between our approach and some existing state-of-the-art methods on the ACDC-night-val set.

moving the CDC and relying only on CDS for knowledge distillation lead to a drop of 2.2% mIoU. We further assess the role of each component in the CDC. Training without the project head astonishingly deteriorates the mIoU by 9.3%, which verifies that the difference in feature distribution caused by the domain gap in the task will seriously affect the effectiveness of knowledge distillation. Features from different domains need to be mapped to the common embedding space to approximate the distribution range, and then the effective correlation knowledge can be extracted. On the basis of the project head, L_{JS} is conducive to further obtaining the content embedding. Experiment shows that disabling the L_{JS} causes a 2.9% mIoU decrease. Subsequently, Cor_{illu} and Cor_{in} will be used in the CDC to realize the distillation of the content correlation across domains, which contribute 1.8% and 1.6% mIoU respectively. Note that after removing the project head or L_{JS} in the CDC, the performance is worse than disabling the CDC completely. This proves from the side that CDS has realized the satisfactory style transfer which is beneficial for the nighttime semantic segmentation. More importantly, it reflects that knowledge distillation in the domain adaptation is very sensitive that the failure to extract the appropriate embedding will be detrimental to the model. In summary, these model variants verify that for domain adaptation with large domain shifts, the adequate and effective use of correlation knowledge within a similar range of data distribution can greatly improve the performance of the model.

The same illumination style between the S_n and T_n is the prerequisite for the cross-domain content distillation. We disable the CDS resulting in a drop of 3.5% mIoU, which is in line with expectations. The semantic-level style alignment implemented by CDS can generate nighttime images aiming at the nighttime semantic segmentation, and obtain the synthetic domain that satisfies our hypothetical domain shift. The LAB-based translation advances the performance about 4%, which reflects that the holistic style alignment can reduce the difficulty of subsequent semantic-level transfer. This also proves that it is not appropriate to employ distillation loss directly when the domain shift is large.

After removing CDC and CDS, the model achieves 44.8% mIoU. The CDS brings a gain of 0.5% mIoU, while only adding CDC, mIoU has dropped by 0.8%. This further illustrates the importance of CDS to achieve semantic-level style transfer, and on the basis of CDS, CDC can further achieve a huge improvement.

5. Conclusions

In this paper, we propose an unsupervised domain adaptation framework via the invariance of cross-domain difference for nighttime semantic segmentation. We validate the effectiveness of properly handling these two kind of domain shifts, *i.e.* illumination and inherent difference. The proposed cross-domain content and style distillation, by extracting the content and style knowledge contained in the features, utilize the invariance of inherent and illumination difference across domains, and realize knowledge distillation and sementic-level style transfer simultaneously. Experiment results verify the effectiveness of our proposed method. Since the distillation is based on the domain shift between source and target domain, it cannot always be effective enough for all nighttime style, which will be further explored in our future work.

Acknowledgement This work is supported by the National Nature Science Foundation of China (No.62171315).

References

[1] 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. 6, 7

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[3] Shuaijun Chen, Xu Jia, Jianzhong He, Yongjie Shi, and Jianzhuang Liu. Semi-supervised domain adaptation based on dual-level domain mixing for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11018–11027, 2021. 2

[4] Xu Chen, Bryan M Williams, Srinivasa R Vallabhaneni, Gabriela Czanner, Rachel Williams, and Yalin Zheng. Learning active contour models for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11632–11640, 2019. 1

[5] Jiaxin Cheng, Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Prem Natarajan. Style-aware normalized loss for improving arbitrary style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 134–143, 2021. 5

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 5

[7] 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. 1, 3, 6, 7

[8] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2021. 3, 4

[9] Qi Dou, Quande Liu, Pheng Ann Heng, and Ben Glocker. Unpaired multi-modal segmentation via knowledge distillation. *IEEE transactions on medical imaging*, 39(7):2415–2425, 2020. 2

[10] Zhekai Du, Jingjing Li, Hongzu Su, Lei Zhu, and Ke Lu. Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3937–3946, 2021. 2

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 4, 5

[12] A Geiger, P Lenz, and R Urtasun. Are we ready for autonomous driving? *The kitti vision benchmark suite,"in IEEE Confer-ence on Computer Vision and Pattern Recognition (CVPR)*, 2011. 1

[13] Jianzhong He, Xu Jia, Shuaijun Chen, and Jianzhuang Liu. Multi-source domain adaptation with collaborative learning for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11008–11017, 2021. 1, 2, 5

[14] 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. 6, 7

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[16] 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. 2

[17] Yuenan Hou, Zheng Ma, Chunxiao Liu, Tak-Wai Hui, and Chen Change Loy. Inter-region affinity distillation for road marking segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12486–12495, 2020. 3

[18] Yunzhong Hou and Liang Zheng. Visualizing adapted knowledge in domain transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13824–13833, 2021. 2

[19] Joy Hsu, Wah Chiu, and Serena Yeung. Darcnn: Domain adaptive region-based convolutional neural network for unsupervised instance segmentation in biomedical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1003–1012, 2021. 2

[20] Hanzhe Hu, Shuai Bai, Aoxue Li, Jinshi Cui, and Liwei Wang. Dense relation distillation with context-aware aggregation for few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10185–10194, 2021. 2, 3

[21] Takashi Isobe, Xu Jia, Shuaijun Chen, Jianzhong He, Yongjie Shi, Jianzhuang Liu, Huchuan Lu, and Shengjin Wang. Multi-target domain adaptation with collaborative consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8187–8196, 2021. 2

[22] Kang Li, Lequan Yu, Shujun Wang, and Pheng-Ann Heng. Towards cross-modality medical image segmentation with online mutual knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 775–783, 2020. 2

[23] Shuang Li, Mixue Xie, Kaixiong Gong, Chi Harold Liu, Yulin Wang, and Wei Li. Transferable semantic augmentation for domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11516–11525, 2021. 2

[24] Xin Li, Deng-Ping Fan, Fan Yang, Ao Luo, Hong Cheng, and Zicheng Liu. Probabilistic model distillation for seman-

tic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7505–7514, 2021. 2, 3

[25] 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. 6, 7

[26] Jian Liang, Dapeng Hu, and Jiashi Feng. Domain adaptation with auxiliary target domain-oriented classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16632–16642, 2021. 2

[27] 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. 4, 6, 7

[28] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen. Structured knowledge distillation for dense prediction. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 3

[29] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2021. 2

[30] Pauline Luc, Natalia Neverova, Camille Couprie, Jakob Verbeek, and Yann LeCun. Predicting deeper into the future of semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 648–657, 2017. 1

[31] 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. 2, 5

[32] Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1094–1103, 2021. 2

[33] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 2

[34] Eduardo Romera, Luis M Bergasa, Kailun Yang, Jose M Alvarez, and Rafael Barea. Bridging the day and night domain gap for semantic segmentation. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1312–1318. IEEE, 2019. 1, 3

[35] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7374–7383, 2019. 1, 2, 3, 5, 6, 7

[36] 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. 1, 3, 5, 6, 7

[37] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. *arXiv preprint arXiv:2104.13395*, 2021. 2, 6

[38] 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. 1, 3

[39] Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1522–1531, 2021. 2, 3, 4

[40] 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. 6, 7

[41] 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. 6, 7

[42] Johan Vertens, Jannik Zürn, and Wolfram Burgard. Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8461–8468. IEEE, 2020. 3

[43] 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. 6, 7

[44] Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, and Kuk-Jin Yoon. Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 608–619, 2021. 2, 3, 4

[45] Lichen Wang, Jiaxiang Wu, Shao-Lun Huang, Lizhong Zheng, Xiangxiang Xu, Lin Zhang, and Junzhou Huang. An efficient approach to informative feature extraction from multimodal data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5281–5288, 2019. 3

[46] Pei Wang, Yijun Li, and Nuno Vasconcelos. Rethinking and improving the robustness of image style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 124–133, 2021. 5

[47] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image

pixel contrast for semantic segmentation. *arXiv preprint arXiv:2101.11939*, 2021. 2, 4

[48] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15769–15778, 2021. 1, 3, 5, 6, 7

[49] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 518–534, 2018. 2

[50] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018. 3

[51] Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. Hierarchical self-supervised augmented knowledge distillation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1217–1223, 2021. 2

[52] 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. 7

[53] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13834–13844, 2021. 2

[54] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 2

[55] 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. 5