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

Huan Gao¹ Jichang Guo¹* Guoli Wang² Qian Zhang²
¹School of Electrical and Information Engineering, Tianjin University.
²Horizon Robotics.
{gh99, jguo}@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.

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
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
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 $T_d$ 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 $P_{T_n}$ 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 $F_D$ 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 $e_D$ from different domains to represent the degree of content difference.

- **Cross-domain style distillation.** Different from the content knowledge, style embedding $G_D$ is obtained by calculating the Gram matrix [11] of the feature $F_D$ 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 $F_D$ into the common embedding space, and get $e_D$. Then we further introduce the *Jensen-Shannon (JS) divergence* to constrain the feature distribution. Specifically,

$$L_{JS} = \lambda (JS(e_{S_d}||e_{S_n}) + JS(e_{T_d}||e_{T_n})) - (JS(e_{S_d}||e_{T_d}) + JS(e_{S_n}||e_{T_n}))$$ (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). $\lambda$ is the coefficient used to control the effect of reverse JS divergence and it is set to 4.

After getting the content embedding $e_p$, 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:

$$Cor_{illu} = \cos(e_{kd}, e_{kn}), k \in \{ S, T \}$$ (2)

$$Cor_{in} = \cos(e_{Sn}, e_{Tn}), r \in \{ d, n \}$$ (3)

$Cor_{illu}$ indicates the correlation of the content within the domain $S$ or the domain $T$, and $Cor_{in}$ indicates the inherent correlation between different datasets in the daytime or nighttime scene. $\cos(x, y) = \frac{x^T y}{\|x\|_2 \|y\|_2}$ 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, $Cor_{illu}$ is used to guide $Cor_{illu}$, and $Cor_{in}$ is used to guide $Cor_{in}$. 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:

$$L_{C_{DC}} = ||\text{Cor}_{illu} - \text{Cor}_{illu_T}||_2^2 + ||\text{Cor}_{in} - \text{Cor}_{in_T}||_2^2 + L_{JS}$$

(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_{illu} 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} reflects the similarity between $S_n$ and $T_n$, and they also have only the inherent difference. Cor_{illu} and Cor_{in} 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_{illu} and Cor_{in} 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 $G_D \in \mathbb{R}^{C \times C}$ to represent the style of the feature, where $G_D$ is the inner product of the vectorised feature maps of $F_D$ on channel i and j respectively:

$$G_D = \sum_p F^{ij}_D F^{jp}_D, D \in \{S_d, S_n, T_d, T_n\}$$

(5)

where p is the pixel of $F_D$. After obtaining the style knowledge of the feature $F_D$ 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:

$$Cor_{G_k} = \cos(G_{k_s}, G_{k_n}), k \in \{S, T\}$$

(6)

$$L_{C_{D_{S}}} = ||Cor_{G_S} - Cor_{G_T}||_2^2$$

(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:

$$L = L_{\text{seg}_n} + L_{\text{seg}_d} + L_{\text{pseudo}} + \lambda_1 L_{C_{DC}} + \lambda_2 L_{C_{D_{S}}}$$

(8)

where $L_{\text{seg}_n}$ is the weighted cross-entropy loss between the prediction map $P_{S_n}$ and the corresponding ground truth, and $L_{\text{seg}_d}$ is in the same way. $L_{\text{pseudo}}$ 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 $\lambda_1$, $\lambda_2$ 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>road</th>
<th>sidewalk</th>
<th>building</th>
<th>wall</th>
<th>fence</th>
<th>traffic light</th>
<th>traffic sign</th>
<th>vegetation</th>
<th>terrain</th>
<th>person</th>
<th>rider</th>
<th>car</th>
<th>truck</th>
<th>bus</th>
<th>train</th>
<th>motorcycle</th>
<th>bicycle</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepLab-v2-Cityscapes [1]</td>
<td>79.0</td>
<td>21.8</td>
<td>53.0</td>
<td>13.3</td>
<td>11.2</td>
<td>22.5</td>
<td>43.5</td>
<td>10.4</td>
<td>18.0</td>
<td>37.4</td>
<td>33.8</td>
<td>64.1</td>
<td>6.4</td>
<td>0.0</td>
<td>52.3</td>
<td>30.4</td>
<td>7.4</td>
<td>28.8</td>
</tr>
<tr>
<td>RefineNet-Cityscapes [27]</td>
<td>68.8</td>
<td>23.2</td>
<td>46.8</td>
<td>20.8</td>
<td>12.6</td>
<td>29.8</td>
<td>30.4</td>
<td>26.9</td>
<td>43.1</td>
<td>14.3</td>
<td>0.3</td>
<td>36.9</td>
<td>49.7</td>
<td>63.6</td>
<td>6.8</td>
<td>0.2</td>
<td>24.0</td>
<td>33.6</td>
</tr>
<tr>
<td>AdaptSegNet-Cityscapes→DZ-night [41]</td>
<td>86.1</td>
<td>44.2</td>
<td>55.1</td>
<td>22.2</td>
<td>4.8</td>
<td>21.1</td>
<td>37.2</td>
<td>8.4</td>
<td>1.2</td>
<td>35.9</td>
<td>67.6</td>
<td>26.7</td>
<td>45.1</td>
<td>50.1</td>
<td>53.9</td>
<td>15.6</td>
<td>30.4</td>
<td></td>
</tr>
<tr>
<td>ADVENT-Cityscapes→DZ-night [43]</td>
<td>85.8</td>
<td>37.9</td>
<td>55.5</td>
<td>14.5</td>
<td>21.4</td>
<td>31.2</td>
<td>14.0</td>
<td>2.0</td>
<td>39.9</td>
<td>16.6</td>
<td>64.0</td>
<td>13.8</td>
<td>58.8</td>
<td>28.5</td>
<td>20.7</td>
<td>29.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BDL-Cityscapes→DZ-night [25]</td>
<td>85.3</td>
<td>41.1</td>
<td>61.9</td>
<td>32.7</td>
<td>17.4</td>
<td>21.3</td>
<td>39.4</td>
<td>1.1</td>
<td>37.4</td>
<td>22.1</td>
<td>63.2</td>
<td>28.2</td>
<td>47.7</td>
<td>39.4</td>
<td>15.7</td>
<td>30.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UDAclustering-Cityscapes→DZ-night [40]</td>
<td>85.5</td>
<td>40.9</td>
<td>59.2</td>
<td>31.2</td>
<td>19.5</td>
<td>24.0</td>
<td>39.1</td>
<td>18.4</td>
<td>39.1</td>
<td>49.7</td>
<td>61.5</td>
<td>34.9</td>
<td>25.8</td>
<td>33.6</td>
<td>19.0</td>
<td>33.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCMA [35]</td>
<td>81.7</td>
<td>46.9</td>
<td>58.8</td>
<td>22.0</td>
<td>20.0</td>
<td>41.2</td>
<td>64.8</td>
<td>31.0</td>
<td>32.1</td>
<td>53.5</td>
<td>47.5</td>
<td>75.5</td>
<td>39.2</td>
<td>49.6</td>
<td>30.7</td>
<td>42.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCMA [35]</td>
<td>80.3</td>
<td>49.3</td>
<td>66.2</td>
<td>7.8</td>
<td>11.0</td>
<td>41.4</td>
<td>38.9</td>
<td>39.0</td>
<td>64.1</td>
<td>18.0</td>
<td>55.8</td>
<td>52.1</td>
<td>53.8</td>
<td>74.7</td>
<td>66.0</td>
<td>42.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DANNet(RefineNet) [48]</td>
<td>90.0</td>
<td>54.0</td>
<td>74.8</td>
<td>41.0</td>
<td>21.1</td>
<td>72.0</td>
<td>26.2</td>
<td>84.0</td>
<td>47.0</td>
<td>33.9</td>
<td>68.2</td>
<td>19.0</td>
<td>66.4</td>
<td>38.3</td>
<td>23.6</td>
<td>44.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>89.6</td>
<td>58.1</td>
<td>70.6</td>
<td>36.6</td>
<td>22.5</td>
<td>33.0</td>
<td>27.0</td>
<td>30.5</td>
<td>68.3</td>
<td>33.0</td>
<td>80.9</td>
<td>42.3</td>
<td>40.1</td>
<td>59.4</td>
<td>58.1</td>
<td>0.1</td>
<td>72.6</td>
<td>47.7</td>
</tr>
</tbody>
</table>

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 $5 \times 10^{-4}$, and the initial learning rate is set as $2.5 \times 10^{-4}$. 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).

<table>
<thead>
<tr>
<th>Method</th>
<th>mIoU1</th>
<th>mIoU2</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepLab-v2-Cityscapes [1]</td>
<td>16.3</td>
<td>17.3</td>
</tr>
<tr>
<td>RefineNet-Cityscapes [27]</td>
<td>20.3</td>
<td>20.4</td>
</tr>
<tr>
<td>AdaptSegNet-Cityscapes→DZ-night [41]</td>
<td>23.8</td>
<td>22.0</td>
</tr>
<tr>
<td>ADVENT-Cityscapes→DZ-night [43]</td>
<td>26.2</td>
<td>22.6</td>
</tr>
<tr>
<td>BDL-Cityscapes→DZ-night [25]</td>
<td>23.9</td>
<td>22.8</td>
</tr>
<tr>
<td>UDAclustering-Cityscapes→DZ-night [40]</td>
<td>24.5</td>
<td>20.0</td>
</tr>
<tr>
<td>GCMA [35]</td>
<td>-</td>
<td>33.2</td>
</tr>
<tr>
<td>MGCDA [36]</td>
<td>29.0</td>
<td><strong>34.9</strong></td>
</tr>
<tr>
<td>DANNet(RefineNet) [48]</td>
<td>37.0</td>
<td>30.3</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>37.7</strong></td>
<td><strong>33.0</strong></td>
</tr>
</tbody>
</table>

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.
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 semantic-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


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


[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


[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


[41] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulte, Kihyuk Sohn, Ming-Hsuan Yang, and Mannoham 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


