2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Towards Dynamic and Small Objects Refinement for Unsupervised Domain Adaptative Nighttime Semantic Segmentation

Jingyi Pan[1], Sihang Li[1], Yucheng Chen[1], Jinjing Zhu[1] and Lin Wang[1,2*]

***Abstract*— Nighttime semantic segmentation plays a crucial role in practical applications, such as autonomous driving, where it frequently encounters difficulties caused by inadequate illumination conditions and the absence of well-annotated datasets. Moreover, semantic segmentation models trained on daytime datasets often face difficulties in generalizing effectively to nighttime conditions. Unsupervised domain adaptation (UDA) has shown the potential to address the challenges and achieved remarkable results for nighttime semantic segmentation. However, existing methods still face limitations in 1) their reliance on style transfer or relighting models, which struggle to generalize to complex nighttime environments, and 2) their ignorance of dynamic and small objects like vehicles and poles, which are difficult to be directly learned from other domains. This paper proposes a novel UDA method that refines both label and feature levels for dynamic and small objects for nighttime semantic segmentation. First, we propose a dynamic and small object refinement module to complement the knowledge of dynamic and small objects from the source domain to target the nighttime domain. These dynamic and small objects are normally context-inconsistent in under-exposed conditions. Then, we design a feature prototype alignment module to reduce the domain gap by deploying contrastive learning between features and prototypes of the same class from different domains, while re-weighting the categories of dynamic and small objects. Extensive experiments on three benchmark datasets demonstrate that our method outperforms prior arts by a large margin for nighttime segmentation. Project page: https://rorisis.github.io/DSRNSS/.**

I. Introduction

Semantic segmentation is essential for scene understanding of autonomous driving [1]–[6] and robotic systems [7]–[10]. Recently, with the development of deep neural networks (DNNs), semantic segmentation has achieved remarkable progress. However, existing methods predominantly target the daytime images [11]–[13]. Thus their performance drastically drops in challenging environments, *e.g.*, nighttime [14]–[16], especially to dynamic and small objects. There are many approaches leveraging other modalities to assist the nighttime segmentation, *i.e.*, thermal cameras [17], [18] and event cameras [19], [20], but the bottleneck of single-modal semantic segmentation at nighttime lies in the decreased performance resulting from inadequate illumination and the scarcity of annotated datasets.

To tackle the problem, unsupervised domain adaptation (UDA) methods have been developed to adapt the model learning unlabeled target domain (*i.e.*, nighttime) images from the model trained with the source domain (*i.e.*, daytime) images. The prevailing methods can be categorized into three types: 1) Some methods, *e.g.*, [14], [15], [21]–[24], use the style transfer models, *e.g.*, CycleGAN [25], to generate daytime or nighttime images, which serve as an intermediate domain to connect the source and target domains. However, these methods are cumbersome as they require multi-stage learning, and the performance can't be guaranteed if style transfer fails. 2) Other methods utilize twilight images with the coarsely aligned day-night image pairs in the target domain, to progressively adapt from the daytime to nighttime domains [14], [15], [26]. 3) The rest methods leverage prior GPS knowledge or static loss to reduce the influence of coarsely aligned day-night image pairs, which improves the quality of pseudo labels [16], [27]–[30]. However, they pay less attention to ***the dynamic and small objects***, *e.g.*, vehicles and poles, in nighttime images, which are particularly challenging to transfer from the daytime domain to the nighttime domain due to their misalignment as shown in Fig. 1. Also, existing UDA methods employing patch-wise or pixel-wise contrastive learning [31], [32] face challenges in ***acquiring effective semantic contexts within patches or pixels in the under-exposed conditions***.

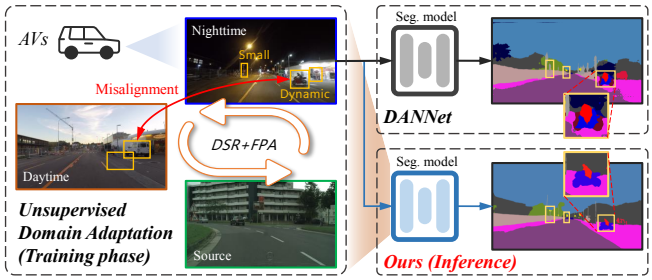


Fig. 1. **Left:** Our UDA method addresses the difficulty of transferring knowledge about dynamic and small objects from other domains for nighttime segmentation. **Right:** Compared with DANNet [16], our method significantly improves the performance of the dynamic and small objects.

Motivation: In this paper, we propose a novel one-stage UDA framework for nighttime semantic segmentation paying more attention to the dynamic and small objects. Specifically, our framework consists of two key technical modules: dynamic and small objects refinement (**DSR**) module and feature prototype alignment (**FPA**) module. Firstly, the DSR module generates a composite mask from the ground truth of the source domain, emphasizing dynamic and small objects, and randomly selected classes. We also introduce a memory bank to store the long-tailed objects across images, which complements these low-occurring categories from different images (Sec. III-B). We observe that the mixup operation forms a new domain—the mixed domain that has a domain shift with the source daytime and target nighttime domain. Therefore, we propose the FPA module to align the source domain with both the mixed and target nighttime domain, aiming to generally reduce the domain gap (Sec. III-C). FPA extracts reliable prototypes from labels in the source domain and pseudo-labels in the mixed domain while pulling the features with the same classes from another domain as positive pairs closer, and pushing other negative pairs away via a contrastive loss. Note that the same operation is conducted in the source and target nighttime domain. We further design an adaptive re-weighting mechanism to improve the attention of some dynamic and small categories in the FPA module. Extensive experiments on

*Corresponding author.
[1]J. PAN, S. Li, Y. Chen, J. Zhu are with the AI Thrust, HKUST(GZ), Guangdong, 511458, China. Email: {jpan305, sli886, ychen208, jzhu706}@connect.hkust-gz.edu.cn
[1,2]L. Wang is with AI/CMA Thrust, HKUST(GZ) and Dept. of CSE, HKUST, Hong Kong SAR, China, Email: linwang@ust.hk
This work was supported by the Guangzhou Fundamental and Applied Basic Research under Grant No.2024A04J4072.

the Dark Zurich [14], the Nighttime Driving [26], and ACDC [33] datasets outperform the state-of-the-art (SoTA) in most categories, particularly in identifying dynamic and small objects. Notably, our approach achieved remarkable mean intersection over union (mIoU) scores of 60.9% for the 'pole' category, 86.8% for the 'car' category, and 25.2% for the 'bus' category, which represents a significant improvement of **2.9%**, **3.0%** and **20.2%** absolute performance, respectively, over the SoTA method.

In summary, the main contributions of this work are three-fold: (**I**) We study a crucial problem of focusing more on improving the performance of dynamic and small objects and reducing the domain shifts caused by illumination and style differences. (**II**) We propose a novel UDA framework for nighttime semantic segmentation framework by designing the dynamic and small objects refinement (DSR) module and feature prototype alignment (FPA) module. (**III**) Extensive experiments on the Dark Zurich, Nighttime Driving, and ACDC-night datasets verify that our method achieves a new SoTA performance for nighttime semantic segmentation.

II. Related work

A. UDA for semantic segmentation

1) Daytime semantic segmentation: Daytime semantic segmentation can be divided into three categories. The first type of approach leverages adversarial learning to reduce the domain gap [4], [34], [35]. The second type of method employs style transfer models, *e.g.*, CycleGAN [25], to build an intermediate domain to bridge the domain gap [36]–[38]. The rest line of the methods impose extra constraints on data distribution to align source and target domains for domain adaptation. [39] utilizes the NCE loss on a vector which is converted from the feature to obtain the global semantic relationship. On the other hand, Liu *et al*. [40] bring the distribution closer to each other by BN statistical information. However, these daytime semantic segmentation models are hard to be applied in nighttime scenarios due to the poor illumination. Thus, we focus on improving the performance of nighttime semantic segmentation, especially for dynamic and small objects.

2) Nighttime semantic segmentation: [14]–[16], [21], [23], [27], [41] utilize style transfer models to enhance the nighttime appearance to resemble daytime lighting. Among them, DANNet [16] designs a relighting network to push the intensity distribution of daytime and nighttime images closer. Bi-Mix [27] proposes a bidirectional mixing framework and exploits day-night image pairs to improve the quality of relighting. However, these methods render the segmentation performance to be highly dependent on the style transfer model. Besides, several works propose to construct an intermediate domain to reduce the domain gap. [14], [15], [26] progressively adapt the daytime-trained model to model learning nighttime images, assisted by the intermediate twilight domain. Also, some methods use other sensors to assist nighttime segmentation [17], [42]. *However, these methods ignore the dynamic and small objects, e.g., vehicles and poles, in the domain alignment process, and the capacity to address the inherent domain shifts between datasets caused by illumination and style differences*.

B. Mixup

It is a strategy to augment data to improve the robustness of DNNs. It has been widely used to augment samples between the source and target domains to reduce the domain gap. [43] [44] [45] exploit the mixup ratio combined with randomly sampled values from the beta distribution to achieve domain adaptation. However, mixup samples are inclined to have local ambiguity and artifacts, which result in model confusion. Cutmix [46] is similar to general mixup while the difference is that it removes pixels and adds removed regions with regions from another image. After that, Walawalkar *et al*. [47] propose attentive cutmix to enhance the cutmix strategy for better generalization. Classmix [48] is a generalization of cutmix, which utilizes a binary mask to mix randomly sampled images. *Unlike previous mixup strategies, we propose a dynamic and small object refinement strategy, which enables the nighttime domain to enhance these weak categories from the source domain, i.e., dynamic and small objects.*

III. Method

A. Overview

An overview of our proposed framework is shown in Fig. 2. The proposed approach involves a labeled source domain S_d and two coarsely aligned, unlabeled target domains T_d and T_n. In the source domain, we have pixel-level annotated source data $X_s \in S_d$ and extract feature representations f_s using the student model F_s. We then obtain segmentation predictions y_s supervised by cross-entropy loss $\mathcal{L}sup$ with Y_s . Furthermore, the target daytime $X_d \in T_d$ and nighttime images $X_n \in T_n$ are unlabeled. We use the teacher network F_t and align the large static regions present in both daytime and nighttime target images through holistic refinement [49]. By replacing the aligned regions in nighttime predictions with those of daytime predictions, we obtain a refined nighttime pseudo label y'_n. We then utilize the Dynamic and Small Objects Refinement module (DSR, Sec. III-B) to obtain extra supervision between X_m and Y_m. Additionally, our proposed Feature Prototype Alignment (FPA, Sec. III-C) module aligns the source domain, mixed domain, and target nighttime domain to effectively reduce domain gaps across various domains. Our aim is to learn and optimize the student model F_s that obtains more knowledge on the dynamic and small objects from labeled source data and unlabeled target daytime domain. It is important to note that the student model and teacher model share the same structure. Detailed descriptions of these modules are provided below.

B. Dynamic and Small Object Refinement

1) Image-level Mixup: Owing to the coarse alignment between target nighttime images and their daytime counterparts, there is a lack of direct supervision for dynamic and small objects in nighttime scenes, making it challenging to obtain accurate pseudo labels from the daytime domain. To address these challenges, we first leverage the labels of the source images to mix up the regions of dynamic and small objects from the source domain into the nighttime images for dynamic and small object refinement (See Fig. 3). By doing so, we can provide accurate predictions on these previously overlooked objects.

Based on the labels Y_s of source images, we define the mixed mask as:

$$M(h,w) = \begin{cases} 1, & \text{if } Y_s(h,w) \in c \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

where M is a binary mask, and c is the selected class. The parameters $h \in H$ and $w \in W$ represent the height and width of the image, respectively.

We further generate a composite mask, denoted as M_c, including the randomly selected class mask M_r and the dynamic and small object mask M_m that contains classes of small and dynamic objects in the source images.

$$M_c = M_r \cup M_m. \tag{2}$$

Based on the composite mask M_c, we augment the nighttime images with complementary objects from the source domain, especially for dynamic and small objects. For image-level mixup,

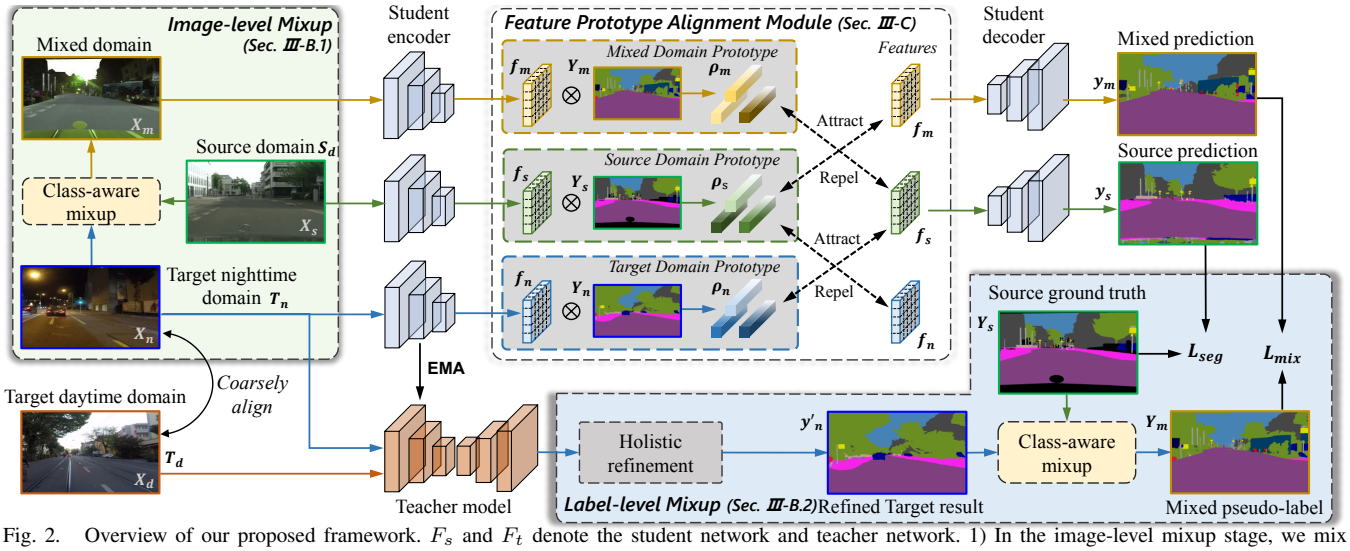


Fig. 2. Overview of our proposed framework. F_s and F_t denote the student network and teacher network. 1) In the image-level mixup stage, we mix dynamic and small classes in source images into target nighttime images according to the ground truth of source images. 2) In the FPA module, we calculate prototypes (ρ_m, ρ_s, ρ_n) of each class to regularize the pixel embedding space (f_s, f_m, f_n) with contrastive learning. 3) In the label-level mixup stage, the refined target results obtained from the holistic refinement and source ground truth are mixed in the same classes as the image-level mixup into the mixed pseudo label for focusing the dynamic and small objects refinement.

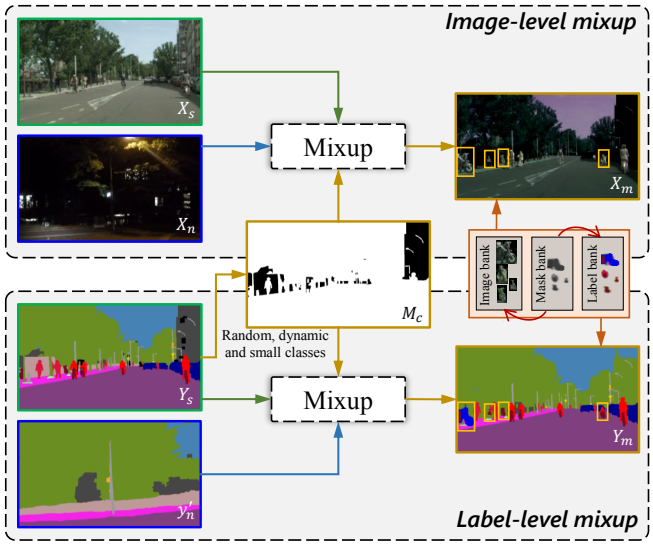


Fig. 3. An illustration of our DSR module. At the image level, two images are obtained from S_d and T_n, respectively. We generate a composite mask from Y_s in the source domain, including randomly selected, dynamic and small classes. This mask is then used to mix the two images while also mixing Y_s and y'_n. Furthermore, the long-tailed memory bank mechanism introduces a few occurrence regions into the mixed image and label.

we utilize the composite mask to mixup corresponding regions of source images and target images as follows:

$$X_m = M_c \odot X_n + (1 - M_c) \odot X_n, \tag{3}$$

here, $\odot$ represents element-wise multiplication.

2) Label-level Mixup: To improve supervision within the mixed image domain, it is essential to generate reliable pseudo labels through label-level alignment facilitated by the composite mask, as outlined in Eq. 2. Thus, we mixup the labels of source and nighttime images corresponding to the image-level mixup as:

$$Y_m = M_c \odot Y_s + (1 - M_c) \odot y'_n, \tag{4}$$

where y'_n is the refined pseudo labels of nighttime images generated using target daytime images.

In addition, addressing the challenge of long-tailed classes, such as buses and bicycles, requires strategies that ensure these classes appear frequently enough to be learned effectively. To this end, we incorporate instances of long-tailed classes from a variety of source images into the target nighttime domain. To facilitate this, we have devised three memory banks: an image bank B_i, a mask bank B_m, and a label bank B_l, which serve to preserve information pertaining to long-tailed classes from different images. These memory banks operate on a first-in-first-out principle, retaining a specified quantity of content at any given time. Next, we mixup the regions of long-tailed classes from the source domain to the above mixed images (X_m, Y_m) as follows:

$$\begin{aligned} X'_m &= B_m \odot B_i + (1 - B_m) \odot X_m, \\ Y'_m &= B_m \odot B_l + (1 - B_m) \odot Y_m. \end{aligned} \tag{5}$$

Furthermore, we utilize the student model F_s to make predictions $y'_m = F_s(X'_m)$ of mixed image X'_m. To ensure consistency between the mixed predictions from F_s and the mixed pseudo labels from F_t, we define a mixup loss as follows:

$$\mathcal{L}_{mix} = -\sum_{c=0}^{C}\sum_{w=0}^{W}\sum_{h=0}^{H} Y'^{c,h,w}_m log(y'^{c,h,w}_m). \tag{6}$$

The weights ϕ'_t of the teacher model during training are updated at every t step by the student's weights ϕ_t:

$$\phi'_t = \lambda\dot{\phi}'_{t-1} + (1 - \lambda)\dot{\phi}_t, \tag{7}$$

where λ is the EMA decay of the smoothing coefficient, and $\lambda \in [0, 1]$.

C. Feature Prototype Alignment

In the context of the DSR module, the mixup operation generates a new domain, referred to as the mixed domain. However, the mixed domain introduces a new domain shift that needs to be addressed to effectively align this domain with the source and target nighttime domains. To overcome this issue, we propose a feature prototype alignment approach to align the mixed domain with the source and target nighttime domain. Note that We also implement alignment of the source domain with the target nighttime domain using a similar approach to reduce the domain gap across the different domains effectively.

To learn domain-invariant features, we aggregate pixel-level features for the same object class across different domains. Firstly, we obtain prototypes of each class by using the ground truth Y_s from the source domain and mixed pseudo label Y'_m from the mixed domain:

$$\rho_s^c = \frac{\sum_h^H \sum_w^W f_s^{h,w} Y_s^{c,h,w}}{\sum_h^H \sum_w^W Y_s^{c,h,w}}, \rho_m^c = \frac{\sum_h^H \sum_w^W f_m^{h,w} Y_m'^{c,h,w}}{\sum_h^H \sum_w^W Y_m'^{c,h,w}}, \tag{8}$$

where ρ_s^c and ρ_m^c represent the prototypes of the source domain and mixed domain for class c, while f_s and f_m denote the source and mixed domain extracted by the student model F_s, respectively. We use these prototypes to compute the cross-domain contrastive loss with the features in another domain. Specifically, we select pixel-wise features and the prototype of the same classes between the source domain and mixed domain as positive pairs, while others are set as negative pairs. This helps us to maximize the pixel-prototype similarity within the same classes:

$$S_{m\to s}^c = (s(f_m^{h,w}, \rho_s^{c,h,w})/\tau) \cdot W^c, \tag{9}$$

where $s(\cdot,\cdot)$ denotes cosine similarity, and τ is a temperature parameter, while W^c is the similarity weight of class c. We can calculate the contrastive loss as follows.

$$\mathcal{L}_{m\to s} = -\sum_c^C \sum_h^H \sum_w^W y_n'^{c,h,w} log \frac{exp(S_{m\to s}^c)}{\sum_c exp(S_{m\to s}^c)}, \tag{10}$$

here, $\mathcal{L}_{m\to s}$ represents the contrastive loss that measures the distance between the source prototypes and mixup features. Additionally, the prototypes of the mixed domain can effectively capture positive pairs belonging to the same class in the source domain while simultaneously repelling negative pairs.

$$S_{s\to m}^c = (s(f_s^{h,w}, \rho_m^{c,h,w})/\tau) \cdot W^c,$$
$$\mathcal{L}_{s\to m} = -\sum_c^C \sum_h^H \sum_w^W Y_s^{c,h,w} log \frac{exp(S_{s\to m}^c)}{\sum_c exp(S_{s\to m}^c)}. \tag{11}$$

To address the issue of class imbalance, we propose an adaptive re-weighting algorithm specifically for cases where the overlapping regions contain dynamic and small objects. By assigning weights to these prototypes, we can adjust the corresponding pixel-wise features accordingly. We set the weights of cosine similarity as:

$$W^c = \begin{cases} 1, & \text{if } c \in C_o \\ s + 1/s & \text{if } c \in C_l \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where W denotes the weight of the contrastive loss, and s represents the number of overlapping classes. The set C_o contains the overlapping classes while not including the long-tailed classes, whereas C_l comprises the long-tailed classes that are also part of the overlapping classes. The total prototype contrastive loss is:

$$\mathcal{L}_{proto} = \mathcal{L}_{n\to s} + \mathcal{L}_{s\to n} + \mathcal{L}_{m\to s} + \mathcal{L}_{s\to m}. \tag{13}$$

D. Objective functions

In addition to the loss functions described in Sec. III-B and Sec. III-C, we also utilize cross-entropy loss to supervise our segmentation model with labeled source images.

$$\mathcal{L}_{sup} = -\sum_{h=0}^H \sum_{w=0}^W \sum_{c=0}^C \mathbb{1}(Y_s^{c,h,w}) log(y_s^{c,h,w}), \tag{14}$$

where $\mathbb{1}(\cdot)$ is the one-hot encoding [53] operation.

The objective of the entire network is formulated as:

$$\mathcal{L} = \mathcal{L}_{sup} + \alpha \mathcal{L}_{mix} + \beta \mathcal{L}_{proto}. \tag{15}$$

where α and β are hyper-parameters to balance three terms. In this work, we set the values as follows: $\alpha = 1.0$, $\beta = 0.1$.

IV. Experiments

A. Datasets

For all experiments, we use the mean of category-wise Intersection-over-union (mIoU) as the evaluation metric. The following datasets are utilized for model training and performance evaluation:

Cityscapes [54] This dataset contains 5,000 frames with pixel-level annotations of 19 categories from street scenes in 50 cities. The images have a resolution of 2,048 x 1,024 pixels. It consists of 2,975 training images, 500 validation images, and 1,525 testing images.

Dark Zurich [14] The Dark Zurich dataset includes 3,041 daytime images, 2,920 twilight images, and 2617 nighttime images captured with coarse aligned in three conditions. We utilize 2,416 day-night image pairs for training, 151 for testing, and 50 for validation.

The input for our source domain comes from Cityscapes [54], while the input of our target domain consists of daytime and nighttime images from Dark Zurich [14]. Additionally, we evaluate our nighttime segmentation model on Nighttime Driving test set [26] and ACDC-night val set [33].

B. Implementation details

We implement our method by using Pytorch-lightning on single Nvidia RTX A6000 GPU. We choose HRDA [52], DeepLabv2 [35] and RefineNet [50] as our based methods. Both our models are trained by the AdamW [55] optimizer with a weight decay of 1×10^{-2}, and the initial learning rate is set as 6×10^{-4}, and linear learning rate warmup with warmup ratio 10^{-6}. Then the learning rate is decreased with the poly learning rate policy with a power of 0.9. The batch size is set to 2. The total training iterations are set to 40k. An EMA factor is $\alpha = 0.999$. At the inference time, there is no change to be introduced to the final model F_s. The training resolution follows the used UDA methods (half resolution for DeepLabv2, and RefineNet and full resolution for HRDA).

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

Comparison on Dark Zurich. We compare our proposed method with some existing state-of-the-art methods, incuding GCMA [14], MGCDA [15], DANNet [16], Bi-Mix [27], CCDistill [24], and several other domain adaptation methods [49], [51], [52]. In order to maintain a fair comparison, we employ a consistent backbone when comparing each method and present the results of corresponding source-only models. The outcomes are reported in Tab. I.

As seen in Tab. I, our method delivers an increase in mIoU of approximately **1.3%**, surpassing the state-of-the-art method Refign (HRDA) [49]. Furthermore, our method demonstrates significant performance improvements over other methods in several categories, particularly in misaligned regions such as vegetation, sky, pole, car, truck, bus, train, motorcycle, and bicycle, where bus and bicycle are categories with relatively fewer occurrences. While scores for other categories differ slightly from the best scores, we achieve competitive performance in each backbone. Our methods exhibit significant improvements over other techniques, particularly in detecting dynamic and small objects. Although there is a degree of randomness in the scores across various categories, the results of our method with three different backbones demonstrate clear advantages in identifying the majority of dynamic and small objects.

Furthermore, we present qualitative results on Dark Zurich-val in Fig. 4, which corroborate our quantitative results. The qualitative

TABLE I

COMPARISON WITH THE STATE-OF-THE-ART METHODS AND BASELINE MODELS ON THE DARK ZURICH-TEST SET. THE BEST RESULTS ARE PRESENTED IN **BOLD**.

Method	road	sidewalk	building	wall	fence	vegetation	terrain	sky	pole	light	sign	person	rider	car	truck	bus	train	motocycle	bicycle	mIoU
	Large static objects								Dynamic and small objects											
RefineNet [50]-Cityscapes	68.8	23.2	46.8	20.8	12.6	43.1	14.3	0.3	29.8	30.4	26.9	36.9	49.7	63.6	6.8	0.2	24.0	33.6	9.3	28.5
GCMA [14]	81.7	46.9	58.8	22.0	20.0	64.8	31.0	32.1	41.2	40.5	41.6	**53.5**	47.5	75.5	39.2	0.0	49.6	30.7	21.0	42.0
MGCDA [15]	80.3	49.3	66.2	7.8	11.0	64.1	18.0	55.8	41.4	38.9	39.0	52.1	**53.5**	74.7	66.0	0.0	37.5	29.1	22.7	42.5
DANNet [16]-RefineNet	90.0	54.0	74.8	41.0	21.1	72.0	26.2	84.0	25.0	26.8	30.2	47.0	33.9	68.2	19.0	0.3	66.4	38.3	23.6	44.3
Bi-Mix [27]	89.2	59.4	75.8	41.7	19.2	70.9	30.1	81.9	39.0	31.9	31.5	44.9	41.8	66.3	34.2	1.0	61.1	47.4	14.6	46.5
CCDistill [24]	89.6	58.1	70.6	36.6	**22.5**	68.3	**33.0**	80.9	33.0	27.0	30.5	42.3	40.1	69.4	58.1	0.1	**72.6**	**47.7**	21.3	47.5
Ours (RefineNet)	**93.9**	**68.2**	**77.4**	**32.3**	11.9	**72.9**	32.4	**85.7**	**50.4**	**45.4**	**44.2**	42.7	30.4	**75.8**	**92.9**	**2.5**	59.8	15.6	**28.4**	**50.7**
DeepLab-v2 [35]-Cityscapes	79.0	21.8	53.0	13.3	11.2	43.5	10.4	18.0	22.5	20.2	22.1	37.4	33.8	64.1	6.4	0.0	52.3	30.4	7.4	28.8
DANNet [16]-DeepLab-v2	88.6	53.4	69.8	34.0	**20.0**	**69.5**	32.2	**82.3**	25.0	31.4	35.9	44.2	43.7	54.1	22.0	0.1	40.9	36.0	24.1	42.5
SePiCo [51]	91.2	61.3	67.0	28.5	15.5	65.4	22.5	80.4	44.7	44.3	**41.3**	41.3	**52.4**	71.2	39.3	0.0	39.6	27.5	28.8	45.4
Ours (DeepLab-v2)	**94.3**	**68.9**	**77.4**	**37.4**	17.5	67.6	**35.4**	**82.3**	**46.6**	**45.6**	41.2	**53.7**	47.7	**79.6**	**71.3**	**6.7**	**71.7**	**40.8**	**34.5**	**53.7**
HRDA [52]-Cityscapes	90.4	56.3	72.0	39.5	19.5	59.3	29.1	70.5	57.8	52.7	43.1	60.0	**58.6**	84.0	75.5	11.2	90.5	**51.6**	40.9	55.9
Refign [49]-HRDA	**95.4**	**76.1**	**86.9**	**53.7**	**37.1**	79.4	**42.0**	91.3	58.0	**57.2**	**63.7**	**63.8**	56.6	83.8	**85.6**	5.0	91.7	47.4	39.5	63.9
Ours (HRDA)	95.2	75.1	85.0	52.0	36.2	**80.2**	38.8	**91.8**	**60.9**	56.9	63.5	62.5	57.3	**86.8**	**85.6**	**25.2**	**92.1**	**51.6**	**42.1**	**65.2**

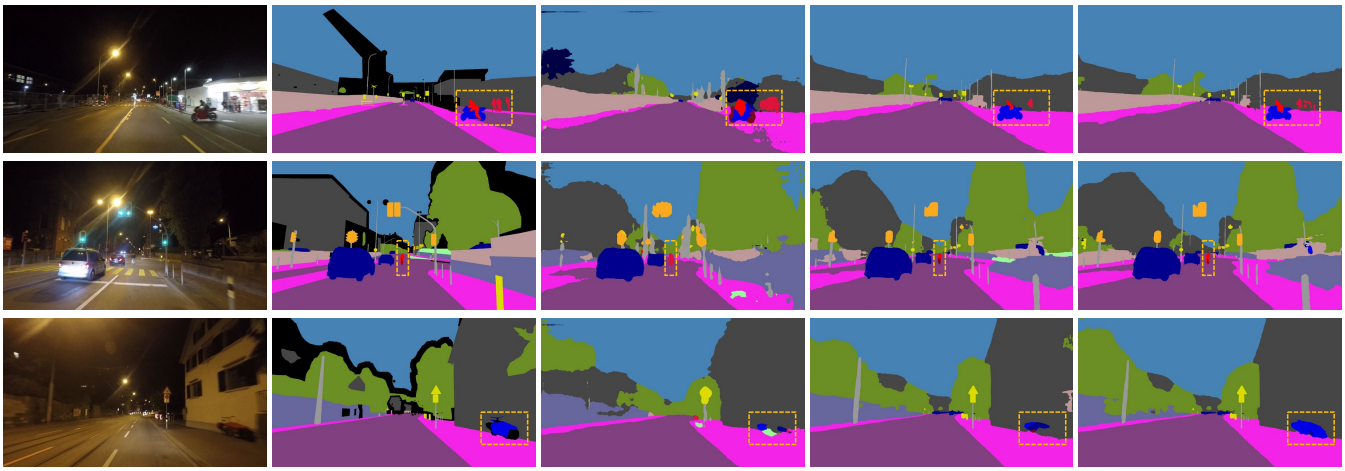


Fig. 4. **Qualitative** comparison between our approach and the state-of-the-art methods on the **Dark Zurich-val set**. Better viewed when zoomed in.

TABLE II

COMPARISON WITH STATE-OF-THE-ART METHODS ON NIGHTTIME DRIVING TEST SETS AND DARK ZURICH-VAL SETS. THE BEST RESULTS ARE PRESENTED IN **BOLD**. 'D', MEANS DEEPLAB-V2 [35], 'R' MEANS REFINENET [50], AND 'T' MEANS HRDA [52].

Method	Backbone	mIoU		
		Nighttime Driving	Dark Zurich-val	ACDC-night
RefineNet [50]-Cityscapes	R	32.8	15.2	20.3
DeepLab-v2 [35]-Cityscapes	D	25.4	12.1	16.3
GCMA [14]	R	45.6	26.7	-
MGCDA [15]	R	49.4	26.1	-
DANNet [16]-RefineNet	R	42.4	30.0	37.0
CCDistill [24]	R	46.2	-	37.3
Refign [49]-HRDA	T	58.0	47.5	53.2
Ours (RefineNet)	R	50.8	34.5	40.3
Ours (DeepLabv2)	D	52.4	37.7	43.6
Ours (HRDA)	T	**58.4**	**48.5**	**53.8**

results reveal that our method can accurately predict intricate details of small and dynamic objects, such as the predictions of a person's legs in the first and second rows, and the motorcycle's details in the third row.

Comparison on other Datasets. To confirm the generality of our approach, we present the performance of our method and compare results with previous methods on the Nighttime driving test set, and ACDC-night val set in Tab. II. Our method with HRDA achieves a **0.4%** and **0.6%** increase in mIoU over the best score in the Nighttime driving test set and ACDC-night val set, respectively, while the performance of our method outperforms other methods with the same backbone. In Fig 5 and Fig. 6, we provide qualitative results of Nightime driving test set and ACDC-night val set that clearly demonstrate the enhanced accuracy of our method in predicting the details of dynamic and small objects. It is worth noting that the Nighttime driving test set has coarse annotation. Despite this limitation, our method still achieves significant performance improvements with different backbones.

D. Ablation study

In this section, we train several model variants for 40k iterations and test them on Dark Zurich-test due to their fine-grained pixel-wise annotation and completed categories.

Effectiveness of each component. we conduct extensive experi-

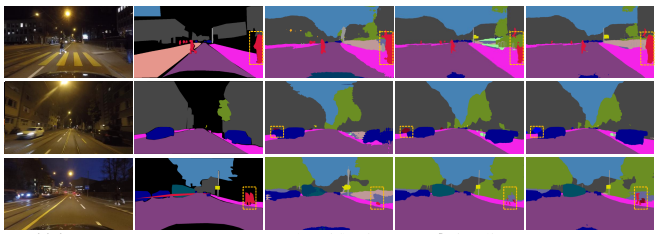


Fig. 5. **Qualitative** comparison between our approach and some existing state-of-the-art methods on the **Nighttime driving test set**. Better viewed when zoomed in.

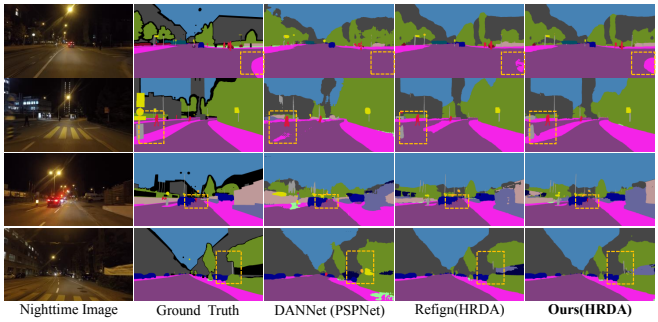


Fig. 6. **Qualitative** comparison between our approach and the state-of-the-art methods on the **ACDC-night val set**. Better viewed when zoomed in.

ments to validate the effectiveness of different components of our method with the same settings and hyperparameters. The results are presented in Tab. III. Adding the composite mask M_c for input and dynamic and small objects refinement improved performance to 63.2 mIoU. Incorporating the long-tailed bank further increased performance by 0.5%. The prototype loss L_{proto} and re-weighting strategy successfully addressed domain shifts and improved alignment between domains.

Effectiveness of the DSR selected classes. To demonstrate the effectiveness of our DSR module, we select different regions in our method and train them with the same implementation for a fair comparison. The results are presented in Tab. IV. Our method, combining randomly selected classes and misaligned regions, achieved a +5.2% mIoU improvement for the baseline and +2.7% mIoU improvement for the randomly selected classes method. This validates the effectiveness of our domain adaptive nighttime semantic segmentation approach.

Effectiveness of the loss weight. We perform a sensitivity analysis on our method's hyperparameters α and β to evaluate their impact on performance, as presented in Tab. V. By varying the values of α and β, we measure their effect on the mIoU. The results indicate that the mIoU remains relatively stable when β remains fixed and α increases significantly. Based on these results, we select the initial weights of $\alpha = 1$ and $\beta = 0.1$.

E. Discussion

1) Refinement of dynamic and small objects: As depicted in Fig. 7, despite the complexity of the environment and the presence of numerous dynamic and small objects, such as people, cars, poles, lights, and bicycles, our segmentation results demonstrate clearer edges and more accurate categorization compared to the baseline. This is the case even in conditions of blur and extremely complex light.

2) All-day performance: Our method is designed to prioritize nighttime image segmentation while ensuring that it remains highly effective for daytime images when compared to established UDA baseline models in Cityscapes-val dataset [54], as shown in Tab. VI.

TABLE III

ABLATION STUDY OF EACH COMPONENT.

Baseline [52]	DSR(w/o long-tailed bank)	DSR(full)	FPA(w/o re-weighting)	FPA(full)	mIoU
✓					55.9
✓	✓				63.2
✓	✓	✓			63.7
✓	✓	✓	✓		64.8
✓	✓	✓	✓	✓	65.2

TABLE IV

ABLATION RESULTS OF SELECTED MIXUP CLASSES IN DSR MODULE.

Method	mIoU	Δ mIoU (%)
Baseline + FPA	60.0	-
+ random classes [48]	62.5	+2.5
+ random classes + small classes	63.7	+3.7
+ random classes + dynamic classes	64.1	+4.1
Ours	**65.2**	**+5.2**

TABLE V

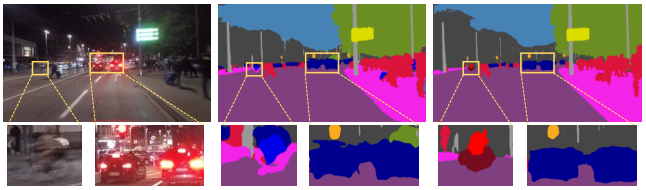
ABLATION RESULTS ON LOSS WEIGHTS. α AND β.

β \ α	0.75	1.00	1.25	1.50
0.075	62.94	63.66	63.41	63.93
0.100	63.99	**65.18**	64.77	64.78
0.125	62.37	64.91	63.21	62.65

TABLE VI

PERFORMANCE ON CITYSCAPES-VAL.

Method	MIoU
DAFormer [56]	68.3
HRDA [52]	73.8
ours (HRDA)	**81.7**



(a) Nighttime images (b) Baseline (c) Ours

Fig. 7. Detailed visualization of our method for refining dynamic and small objects with Dark Zurich-test set [14].

Overall, our experimental results demonstrate the ability of our method to segment all-day images effectively.

V. CONCLUSION

In this paper, we proposed a novel UDA framework for nighttime segmentation by focusing on dynamic and small object refinements and inherent domain shifts. The DSR module combines the random selection of classes with dynamic and small objects while introducing a long-tailed bank to store long-tailed objects to augment the dynamic and small data, which are hard to segment due to poor illumination. The FPA module further constrains the mixup process through feature-level prototype alignment to reduce the domain shift in different domains while improving the weights of dynamic and small objects during the alignment process. Experimental results demonstrated the effectiveness of our method.

REFERENCES

[1] M. Hofmarcher, T. Unterthiner, J. Arjona-Medina, G. Klambauer, S. Hochreiter, and B. Nessler, "Visual scene understanding for autonomous driving using semantic segmentation," *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 285–296, 2019.

[2] M. Siam, M. Gamal, M. Abdel-Razek, S. Yogamani, M. Jagersand, and H. Zhang, "A comparative study of real-time semantic segmentation for autonomous driving," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 587–597.

[3] H. Blum, P.-E. Sarlin, J. Nieto, R. Siegwart, and C. Cadena, "Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving," in *proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 1–10.

[4] J. Liu, X. Guo, B. Li, and Y. Yuan, "Coinet: Adaptive segmentation with co-interactive network for autonomous driving," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4800–4806.

[5] C. Malone, S. Garg, M. Xu, T. Peynot, and M. Milford, "Improving road segmentation in challenging domains using similar place priors," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3555–3562, 2022.

[6] G. Roggiolani, M. Sodano, T. Guadagnino, F. Magistri, J. Behley, and C. Stachniss, "Hierarchical approach for joint semantic, plant instance, and leaf instance segmentation in the agricultural domain," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9601–9607.

[7] M. Kerkech, A. Hafiane, and R. Canals, "Vine disease detection in uav multispectral images using optimized image registration and deep learning segmentation approach," *Computers and Electronics in Agriculture*, vol. 174, p. 105446, 2020.

[8] S. Palazzo, D. C. Guastella, L. Cantelli, P. Spadaro, F. Rundo, G. Muscato, D. Giordano, and C. Spampinato, "Domain adaptation for outdoor robot traversability estimation from rgb data with safety-preserving loss," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 014–10 021.

[9] S. Liu, J. Cheng, L. Liang, H. Bai, and W. Dang, "Light-weight semantic segmentation network for uav remote sensing images," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8287–8296, 2021.

[10] K. Katuwandeniya, S. H. Kiss, L. Shi, and J. V. Miro, "Multi-modal scene-compliant user intention estimation in navigation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1001–1006.

[11] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 289–305.

[12] R. P. Poudel, S. Liwicki, and R. Cipolla, "Fast-scnn: Fast semantic segmentation network," *arXiv preprint arXiv:1902.04502*, 2019.

[13] V. Guizilini, J. Li, R. Ambruș, and A. Gaidon, "Geometric unsupervised domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8537–8547.

[14] C. Sakaridis, D. Dai, and L. V. 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*, 2019, pp. 7374–7383.

[15] C. Sakaridis, D. Dai, and L. Van Gool, "Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[16] X. Wu, Z. Wu, H. Guo, L. Ju, and S. 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*, 2021, pp. 15 769–15 778.

[17] J. Vertens, J. Zürn, and W. 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)*. IEEE, 2020, pp. 8461–8468.

[18] Y. Sun, W. Zuo, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.

[19] J. Zhang, K. Yang, and R. Stiefelhagen, "Issafe: Improving semantic segmentation in accidents by fusing event-based data," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1132–1139.

[20] J. Cao, X. Zheng, Y. Lyu, J. Wang, R. Xu, and L. Wang, "Chasing day and night: Towards robust and efficient all-day object detection guided by an event camera," *arXiv preprint arXiv:2309.09297*, 2023.

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

[22] ——, "Incremental adversarial domain adaptation for continually changing environments," in *2018 IEEE International conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 4489–4495.

[23] L. Sun, K. Wang, K. Yang, and K. Xiang, "See clearer at night: towards robust nighttime semantic segmentation through day-night image conversion," in *Artificial Intelligence and Machine Learning in Defense Applications*, vol. 11169. SPIE, 2019, pp. 77–89.

[24] H. Gao, J. Guo, G. Wang, and Q. Zhang, "Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9913–9923.

[25] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[26] D. Dai and L. Van Gool, "Dark model adaptation: Semantic image segmentation from daytime to nighttime," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 3819–3824.

[27] G. Yang, Z. Zhong, H. Tang, M. Ding, N. Sebe, and E. Ricci, "Bi-mix: Bidirectional mixing for domain adaptive nighttime semantic segmentation," *arXiv preprint arXiv:2111.10339*, 2021.

[28] H. Lee, C. Han, and S.-W. Jung, "Gps-glass: Learning nighttime semantic segmentation using daytime video and gps data," *arXiv preprint arXiv:2207.13297*, 2022.

[29] W. Liu, W. Li, J. Zhu, M. Cui, X. Xie, and L. Zhang, "Improving nighttime driving-scene segmentation via dual image-adaptive learnable filters," *arXiv preprint arXiv:2207.01331*, 2022.

[30] F. Shen, Z. Pataki, A. Gurram, Z. Liu, H. Wang, and A. Knoll, "Loopda: Constructing self-loops to adapt nighttime semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3256–3266.

[31] I. Alonso, A. Sabater, D. Ferstl, L. Montesano, and A. C. Murillo, "Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8219–8228.

[32] W. Liu, D. Ferstl, S. Schulter, L. Zebedin, P. Fua, and C. Leistner, "Domain adaptation for semantic segmentation via patch-wise contrastive learning," *arXiv preprint arXiv:2104.11056*, 2021.

[33] C. Sakaridis, D. Dai, and L. Van Gool, "Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 765–10 775.

[34] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "Fcns in the wild: Pixel-level adversarial and constraint-based adaptation," *arXiv preprint arXiv:1612.02649*, 2016.

[35] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7472–7481.

[36] J. He, X. Jia, S. Chen, and J. Liu, "Multi-source domain adaptation with collaborative learning for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 008–11 017.

[37] H. Ma, X. Lin, Z. Wu, and Y. 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*, 2021, pp. 4051–4060.

[38] T. Isobe, X. Jia, S. Chen, J. He, Y. Shi, J. Liu, H. Lu, and S. Wang, "Multi-target domain adaptation with collaborative consistency learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8187–8196.

[39] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, "Exploring cross-image pixel contrast for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7303–7313.

[40] Y. Liu, W. Zhang, and J. Wang, "Source-free domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1215–1224.

[41] E. Romera, L. M. Bergasa, K. Yang, J. M. Alvarez, and R. Barea, "Bridging the day and night domain gap for semantic segmentation," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1312–1318.

[42] R. Xia, C. Zhao, M. Zheng, Z. Wu, Q. Sun, and Y. Tang, "Cmda: Cross-modality domain adaptation for nighttime semantic segmentation," *arXiv preprint arXiv:2307.15942*, 2023.

[43] X. Mao, Y. Ma, Z. Yang, Y. Chen, and Q. Li, "Virtual mixup training for unsupervised domain adaptation," *arXiv preprint arXiv:1905.04215*, 2019.

[44] M. Xu, J. Zhang, B. Ni, T. Li, C. Wang, Q. Tian, and W. Zhang, "Adversarial domain adaptation with domain mixup," in *Proceedings*

of the AAAI conference on artificial intelligence, vol. 34, no. 04, 2020, pp. 6502–6509.

[45] Y. Wu, D. Inkpen, and A. El-Roby, "Dual mixup regularized learning for adversarial domain adaptation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*. Springer, 2020, pp. 540–555.

[46] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.

[47] D. Walawalkar, Z. Shen, Z. Liu, and M. Savvides, "Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification," *arXiv preprint arXiv:2003.13048*, 2020.

[48] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "Classmix: Segmentation-based data augmentation for semi-supervised learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1369–1378.

[49] D. Bruggemann, C. Sakaridis, P. Truong, and L. Van Gool, "Refign: Align and refine for adaptation of semantic segmentation to adverse conditions," *arXiv preprint arXiv:2207.06825*, 2022.

[50] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[51] B. Xie, S. Li, M. Li, C. H. Liu, G. Huang, and G. Wang, "Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[52] L. Hoyer, D. Dai, and L. Van Gool, "Hrda: Context-aware high-resolution domain-adaptive semantic segmentation," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*. Springer, 2022, pp. 372–391.

[53] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 578–587.

[54] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[55] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[56] L. Hoyer, D. Dai, and L. Van Gool, "Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9924–9935.