Night-time Semantic Segmentation with Unsupervised Learning and Cross Attention

Jian Cheng chengjian113@qq.com
School of Computer Science and Engineering, Southeast University, Nanjing, 211189, Jiangsu, China

Yanfeng Hu huyf@aircas.ac.cn
Aerospace Information Research Institute, Suzhou, 215123, Jiangsu, China

Yu Dai 220212047@seu.edu.cn
School of Computer Science and Engineering, Southeast University, Nanjing, 211189, Jiangsu, China

Xue Qiao xqiao@mail.ie.ac.cn
Aerospace Information Research Institute, Suzhou, 215123, Jiangsu, China

***Li Yao** yao.li@seu.edu.cn
School of Computer Science and Engineering, Southeast University, Nanjing, 211189, Jiangsu, China
Key Laboratory of Computer Network and Information Integration (Southeast University),Ministry of Education,Nanjing,211189,Jiangsu,China

Junyan Yang yjy-2@163.com
School of Architect, Southeast University, Nanjing, 211189, Jiangsu, China

Abstract

In recent years, semantic segmentation has shown very good performance in daytime scenes. But in night-time scenes, semantic segmentation greatly reduces its accuracy. Due to the lack of large-scale night-time semantic segmentation datasets, it is difficult to directly train segmentation models for night-time scenes. Therefore, it becomes important to adapt the daytime scene segmentation model to the night-time scene without directly using the night-time scene segmentation dataset. In this paper, we propose a framework based on unsupervised learning and cross attention. The proposed method fuses supervised daytime scenes and unsupervised night-time scenes, the supervision information in the daytime scene and the texture information specific to the night-time scene are fully utilized, and the model is adapted to both the daytime scene and the night-time scene. Consistency regularization is used to make segmentation model adapt to the complex and changeable night scene texture and illumination. In view of the coarse correspondence of static objects between day and night image pairs in the Dark Zurich dataset, cross attention is proposed to make the model pay more attention to the parts of the night scene which are similar to the daytime scene. Extensive experiments on Dark Zurich and night-time Driving datasets show that our method obtains better performance in night-time semantic segmentation.

Keywords: Semantic Segmentation, Night-time Vision, Unsupervised Learning

1. Introduction

Semantic segmentation combines image classification[Sermanet et al. (2014)], object detection[Redmon and Farhadi (2018)] and image segmentation[Ronneberger et al. (2015); Chen et al. (2015, 2018)]. It divides the image into regional blocks with certain semantic information through a certain method, and identifies the semantic category of each regional block. Finally, a segmented image with pixel-wise semantic annotations is obtained. Currently, semantic segmentation shows good performance in standard scenes (such as daytime scenes with good lighting conditions) and is widely used in fields such as autonomous driving[Chen et al. (2017b); Shelhamer et al. (2017); Wu et al. (2019); Zhou et al. (2018)]. However, since most of the existing semantic segmentation datasets are collected during the day, the models trained with these datasets cannot adapt to complex night-time scenes. Most of the existing works of auto-driving[Janai et al. (2020); Paden et al. (2016); Schwarting et al. (2018)] focus on day scene but ignore the night scenes which is also very important. The light sources of night scenes mainly come from the lighting of street lamps, car lights and buildings, which is very different from the daytime scenes where the main light source comes from sunlight. The special light source at night results in both overexposed and underexposed images at night. Night scenes will also lead to a large number of missing details and textures of objects, blurred edges and other undesirable phenomena.

Early methods use GAN Zhu et al. (2017) to convert images between daytime and night-time domains Sakaridis et al. (2019), and then generated night-time pseudo-segmentation labels for training. However, the images generated by GAN would have artifacts and strange textures, and an additional GAN model with good performance was trained. It is also time-consuming and labor-intensive to generate night scenes. The latest method DANNET Wu et al. (2021) adopts image relighting network to enhance night-time scene lighting, and integrates image relighting network and segmentation network into a unified framework, making the model more efficient. However, DANNET uses coarsely aligned day-night image pairs to generate pseudo-labels, ignoring the difference between daynight image pairs, which can lead to a large number of errors in pseudo-labels for small objects. Moreover, training the image relighting network together with the segmentation network makes training more difficult and the results generated by the image relighting network more difficult to guarantee.

In this work, we proposes a night-time semantic segmentation algorithm based on unsupervised learning and cross attention, which makes use of the supervision information in the daytime scene and the unique texture information of the night-time scene, so that the model can adapt to the daytime scene and the night-time scene at the same time.

In summary, our contributions are as follows:

1. We uses the existing semantic segmentation model as the backbone, fuses the supervised daytime scene and the unsupervised night-time scene, makes use of the supervised information in the daytime scene and the unique texture information of the night-time scene.

2. Consistency regularization is introduced in this paper so that the segmentation model can adapt to the complex and changeable night scene texture and illumination, and at the same time can resist the extra noise caused by low-light enhancement.

3. Due to the coarse correspondence of static objects between day and night image pairs in the Dark Zurich Sakaridis et al. (2019) dataset, cross attention is proposed to make the model pay more attention to the parts of the night scene that are similar to the daytime scene, so that the model can take advantage of high-level semantic information in daytime scene when trained.

2. Related work

2.1. Semantic segmentation

Modern deep learning methods for semantic segmentation are mostly based on fully convolutional networks. Subsequent developments have studied segmentation models from three main aspects: resolution, context, and edges. Work on resolution upscaling includes adjusting the spatial loss induced in classification networks, e.g., using encoder-decoder schemes or dilated convolutions, and maintaining high resolution, such as HRNet Sun et al. (2019). Work which exploits context includes spatial context, such as PSPNet Zhao et al. (2017) and Deeplabv3 Chen et al. (2017a). Deeplabv3 employs atrous spatial pyramid pooling to embed contextual information, which consists of parallel dilated convolutions with different dilation rates. PSPNet designs a pyramid pooling module to collect contextual priors which contains contextual information at different scales. Methods to improve the segmentation quality of edge regions include PointRend Kirillov et al. (2020) and SegFix Yuan et al. (2020).

2.2. Night-time semantic segmentation

Night-time semantic segmentation is crucial for autonomous driving. Segmentation models trained on daytime dataset perform poorly at night-time scene due to the large domain gap. Dai and Gool (2018) proposed a two-step adaptation method by intermediate domain (twilight domain). Sakaridis et al. (2019) use style transfer model to convert the style of night-time images to the style of daytime images. Xu et al. (2021) proposed a curriculum domain adaptation method to smoothly transfer semantic knowledge from daytime to night-time. All of these methods require training the model in multiple stages. To reduce training cost and avoid the problem of training error accumulation, Wu et al. (2021) proposed a one-stage domain adaptation framework called DANNet which jointly trains image relighting network, semantic segmentation network and two discriminators in one stage. This is more efficient and can achieve better performance than other multi-stage method. However, the object correspondence between day and night image pairs sometimes has a large offset. In this case, the effect of static loss will be limited and the information between day-night image pairs will be largely ignored.

3. Methodology

The proposed method involves a source domain S and two target domains T_d and T_n, where S, T_d and T_n represent Cityscapes (daytime)[Cordts et al. (2016)], Dark Zurich-D (daytime)[Sakaridis et al. (2019)], and Dark Zurich-N (night-time), respectively. Note that

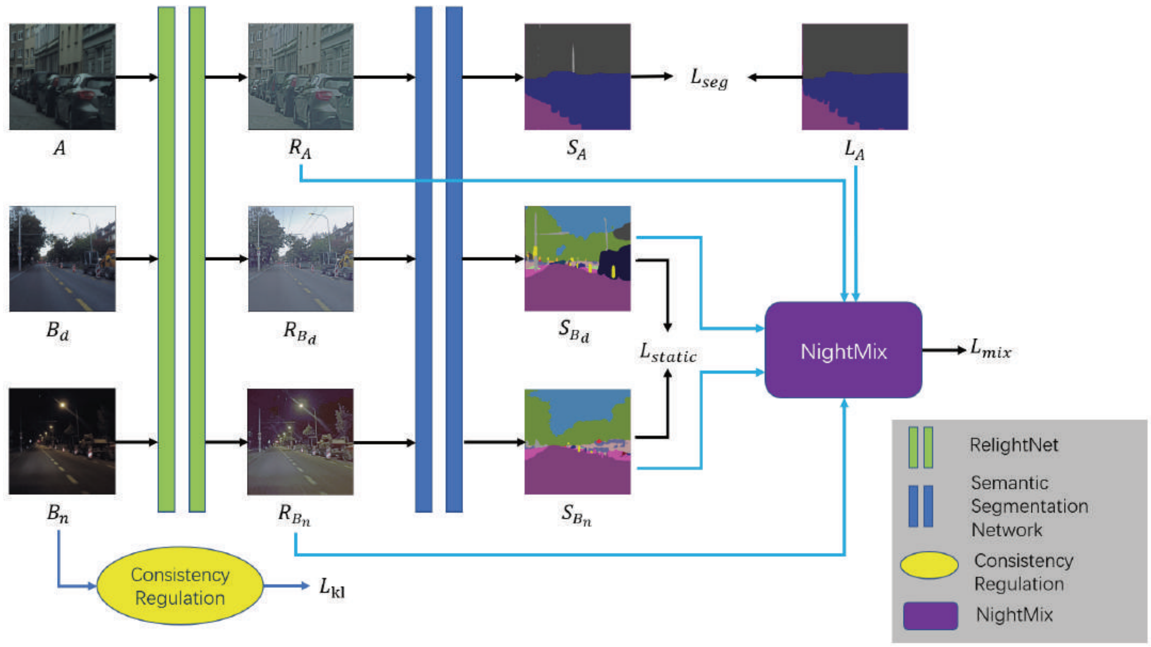


Figure 1: The architecture of the proposed framework. Three input images A, B_d and B_n go through a weight-sharing image relighting network whose weight is fixed. All the outputs are fed into a weight-sharing segmentation network to obtain the predictions. For the predictions from A, a semantic segmentation loss L_{seg} is computed using the ground truth from the source dataset. Besides, the predictions from B_d and B_n are fed into a static loss L_{static} proposed in DANNet. At the same time, R_A, L_A, R_{B_n}, S_{B_n} and S_{B_d} are fed into the NightMix module and finally compute the mixing loss L_{mix}. B_n is fed into the consistency regularization module to compute loss L_{kl}.

only the Cityscapes dataset has ground-truth semantic segmentation in training, and the Dark Zurich dataset has no semantic segmentation labels.

The night-time semantic segmentation framework proposed in this paper is shown in Fig. 1 Images A, B_d and B_n are from source domain S and target domains T_d, T_n. Image A has segmentation ground truth L_A while image B_d and B_n do not have segmentation ground truth. f_θ is Semantic segmentation network, where θ represents the weight to be trained. The image relighting network g is proposed in Zero-DCE Guo et al. (2020). We use the trained model and fix weights in our framework.

3.1. NightMix Module

The NightMix module is modified from the ClassMix algorithm Olsson et al. (2021), as shown in Fig. 2 Image A and B_n are from the source domain S and the target domain T_n, where image A has segmentation ground truth L_A and image B_n has no segmentation ground truth. Image B_n pass through the segmentation network f_θ and output predicted segmentation map S_{B_n}. After remove easy split classes such as sky, road and forest from all classes C_A in L_A, the remaining hard split classes are $\tilde{C}_A$. Half of the hard split classes are randomly selected in $\tilde{C}_A$ and the pixels in these classes are set to 1 in binary mask M, while others are 0. Then we use this mask to mix images A and B_n into the augmented image X_A. Since images B_d and B_n are shot in the same scene, the positions of the static objects (objects that cannot move freely, such as lights, poles, etc.) in two images are relatively close, so we can generate pseudo label $\tilde{S}_{B_n}$ by S_{B_n} and B_d's predicted segmentation map S_{B_d}. The segmentation map L_A and $\tilde{S}_{B_n}$ are also mixed via mask M, generating a mixed pseudo label Y_A of the augmented image X_A.

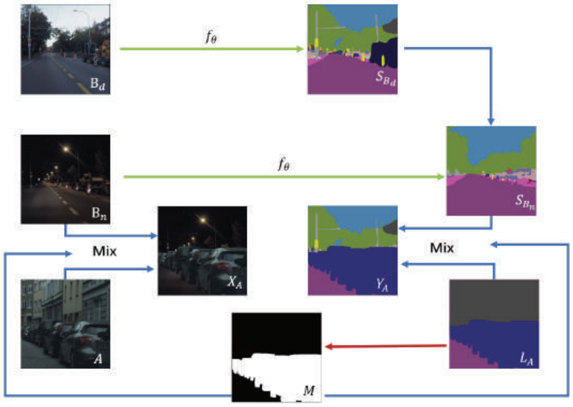


Figure 2: NightMix Module. Image A is from the source domain S, B_n and B_d are from the target domain. Based on the ground truth L_A of image A, a binary mask M is created. The mask is then used to mix the images A and B_n and their respective predictions into an augmented image X_A and the corresponding pseudo label Y_A.

The loss of NightMix module L_{mix} is defined by:

$$L_{mix} = \ell\left(f_\theta\left(X_A\right), Y_A\right) \tag{1}$$

where X_A and Y_A are the augmented images and the corresponding pseudo label, generated by the NightMix module, where the input image A is randomly sampled from the source domain S, and B_n is randomly sampled from the target domain T_n. ℓ is the cross entropy loss.

Since the source domain S has segmentation ground truth, it can be considered that the accuracy of the segmentation model in the daytime scene is much higher than that in the night-time scene. So the night-time predicted segmentation map $\tilde{S}_{B_n}$ corrected by the daytime predicted segmentation map S_{B_d} has a higher accuracy which is sufficient as a pseudo label for night-time scenes. However, if only pseudo label $\tilde{S}_{B_n}$ is used as the label of B_n to train model, the model will tend to easy split classes or classes with a larger size, while ignoring hard split classes, classes with a smaller size and dynamic objects. Therefore, it is necessary to use image A with the segmentation ground truth L_A from the source domain S for NightMix. On the one hand, ground truth can be used to strengthen the cognition of the model, and on the other hand, the model can not only focus on the easy split classes. That's also one of the reasons why we generate $\tilde{C}_A$ from L_A.

3.2. Cross Attention

The night-time images in the Dark Zurich dataset have no ground truth, and there is no pixel-level correspondence between day-night image pairs. But there are coarse correspondences of static objects between day-night image pairs, while dynamic objects have no correspondence at all. In order to make full use of those coarse correspondence, we propose cross attention module, which aggregates the features of day-night image pairs to strengthen similar semantic features, thereby improving the performance of night-time semantic segmentation.

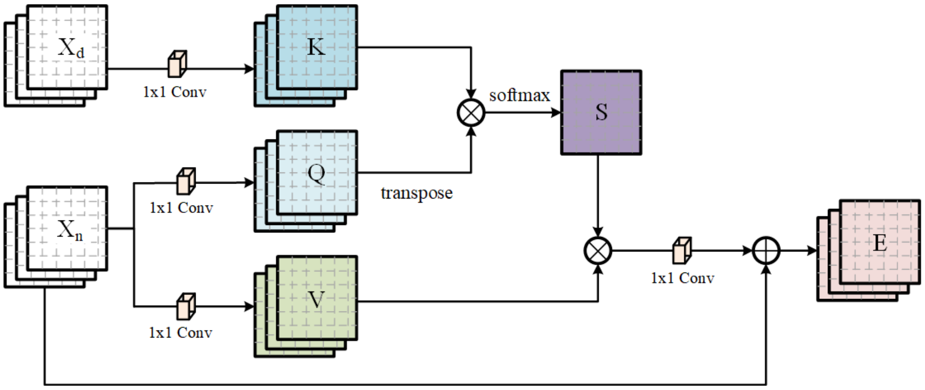


Figure 3: Cross Attention. X_d is the local feature of daytime image B_d and X_n is the local feature of the night-time image B_n corresponding to B_d. E is the final attention-weighted feature map.

As shown in Fig. 3, X_n is a local feature of the night-time image B_n, $X_n \in \mathbb{R}^{C\times H\times W}$. X_d is the local feature of daytime image B_d corresponding to the night-time image B_n, $X_d \in \mathbb{R}^{C\times H\times W}$. X_n and X_d are fed into a convolution layers to generate feature maps Q and K, where $\{Q, K\} \in \mathbb{R}^{C\times H\times W}$. Then we transpose Q and perform matrix multiplication with K, and apply a softmax layer to compute the attention map $S \in \mathbb{R}^{N\times N}$:

$$s_{ji} = \frac{\exp\left(K_i \cdot Q_j\right)}{\sum_{i=1}^{N} \exp\left(K_i \cdot Q_j\right)} \tag{2}$$

where $N = H \times W$ is the number of pixels in the image, where s_{ji} represents the influence of the i^{th} position in the feature map of the daytime image on the j^{th} position in the feature map of the night-time image. The more similar the feature representations of two locations, the greater the correlation between them. At the same time, we feed the feature X_n into a convolutional layer to generate a new feature map $V \in \mathbb{R}^{C\times H\times W}$ and change its shape to $\mathbb{R}^{C\times N}$. Then, we perform a matrix multiplication between the transposed matrix of V and S. Finally, the product is multiplied by a scale parameter α to form an attention-weighted

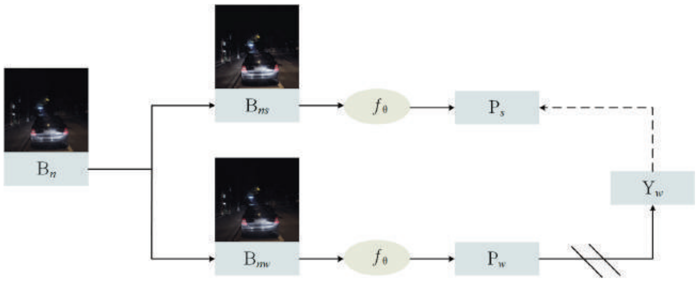


Figure 4: Consistency Regularization. Using pseudo segmentation of weakly augmented images to supervise the predicted segmentation map of strongly augmented images.

feature map, which is added to the original feature map X_n:

$$E_j = \alpha \sum_{i=1}^{N} (s_{ji} V_i) + X_{n_j} \tag{3}$$

Following Fu et al. (2019), α is initialized to 0, is set as a learnable weight for adaptive adjustment during training. Therefore, the final cross-attention feature map E is the weighted sum of the features from all locations in the daytime features and the original features of the night-time image. The computing process of cross attention is modified according to the self-attention mechanism[Hu et al. (2020)].

The cross-attention module models the influence of daytime image features on night-time image features, and strengthens the similar features between them. Because of the difference between night-time images and daytime images, the extracted features of the same class in daytime images and night-time images will be different, and these differences affect the final performance of segmentation. Using cross attention module can establish associations between night-time features and daytime features to adaptively aggregate information from both domains, improving intra-class coherence, thereby improving feature representation for segmentation.

3.3. Consistency Regularization

Consistency regularization[Mittal et al. (2021); Souly et al. (2017)] of the GAN-based method is used in DANNet, forcing the consistency between the statistical features of ground truth (from the source domain S) and the statistical features of the predicted segmentation map of the unlabeled data (from the target domain Tn). However, this method requires an extra discriminator, which increases the complexity of the framework. Moreover, it is found that the training of the discriminator is not stable in our experiments, and it takes multiple trainings and fine-tuning to obtain a stable and effective discriminator.

In this paper, we choose a cheaper and more stable consistency regularization, that is, using pseudo segmentation of weakly augmented images to supervise the predicted segmentation map of strongly augmented images Sohn et al. (2020), as shown in Fig. 4. We generate a strongly augmented image for each unlabeled image, and then compute class probability distribution of the strongly augmented image, and then use it in KL divergence. Given the weakly augmented version B_{nw} of the unlabeled image B_n, we compute the model's predicted class distribution: $P_w = f_\theta(B_{nw})$, terminating the gradient backpropagation at P_w to generate Y_w. Then, a predicted class distribution $P_s = f_\theta(B_{ns})$ is also calculated according to the strongly augmented image B_{ns} of B_n, and finally the KL divergence is used as the distance between the two class probability distributions:

$$L_{kl} = Y_w \log(\frac{Y_w}{P_s}) \tag{4}$$

If the segmentation model generates two different predicted class distributions, the KL divergence will get a large value, which reflects the model's uncertainty about the predictions. In our experiments, weak Gaussian noise is used for weakly augmentation, while Gaussian blur, color shift, brightness contrast shift, and a series of data augmentation methods are used for strongly augmentation. The robustness of the model is increased by consistency regularization, which makes the model perform better and more stable in complex night scenes.

3.4. Loss functions

Due to the imbalance in the number of pixels of different classes in the source domain S, it is easier to converge during model training by predicting pixels as classes with large size, such as roads, buildings, and trees. In this case, it is difficult to correctly make prediction for few and small classes, such as traffic sign and traffic light. To address this issue, we propose a weighting strategy for the predicted class likelihood maps. Specifically, for each class $k \in C$, first define a weight:

$$w_K = 1/(a_k) \tag{5}$$

where a_k is the proportion of all valid pixels in the source domain labeled as class k. Obviously, the smaller the value of a_k, the larger the value of w_k, and using such weights can help segment the classes with smaller size.

We employ a weighted cross-entropy loss to train semantic image segmentation in the source domain:

$$L_{seg} = -\frac{1}{N\|\mathbb{C}\|} \sum_{k\in\mathbb{C}} \left\| w_k GT^{(k)} \cdot \log\left(P_s^{(k)}\right) \right\|_1 \tag{6}$$

where $P_s^{(k)}$ is the k^{th} channel of the predicted P_s from the source image and w_k is the weight defined in Equation 5. $GT^{(k)}$ is the one-hot encoding of the label of the k^{th} class.

If only consider static classes, daytime images have coarse similarities with their corresponding night-time images in Dark Zurich dataset. In this paper, we use the static loss in DANNet Wu et al. (2021) to provide pixel-level pseudo supervision for static classes. Given segmentation predictions $P_{td} \in \mathbb{R}^{H\times W\times C}$ and $P_{tn} \in \mathbb{R}^{H\times W\times C}$, only the channels of static classes are considered to compute static loss. Representing C^S as the total number of static object classes, there are $P_{td}^S \in \mathbb{R}^{H\times W\times C^S}$ and $P_{tn}^S \in \mathbb{R}^{H\times W\times C^S}$. Focal loss Lin et al. (2020) is used to compensate for the imbalance between training samples of different classes. Finally, the static loss L_{static} is defined as:

$$L_{\text{static}} = -\frac{1}{N} \left\| \left(1 - P_{tn}^{\mathcal{S}}\right)^{\gamma} \log(p) \right\|_1 \tag{7}$$

Table 1: The per-category results on Dark Zurich-test by current state-of-the-art methods and our framework. The best results are presented in bold, with the second best results underlined.

Method	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	tarrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	mIoU(%)
AdaptSegNet Valada et al. (2017)	86.1	44.2	55.1	22.2	4.8	21.1	5.6	16.7	37.2	8.4	1.2	35.9	26.7	68.2	45.1	0.0	50.1	33.9	15.6	30.4
ADVENT Vu et al. (2019)	85.8	37.9	55.5	27.7	14.5	23.1	14.0	21.1	32.1	8.7	2.0	39.9	16.6	64.0	13.8	0.0	58.8	28.5	20.7	29.7
DMAda Dai and Gool (2018)	75.5	29.1	48.6	21.3	14.3	34.3	36.8	29.9	49.4	13.8	0.4	43.3	50.2	69.4	18.4	0.0	27.6	34.9	11.9	32.1
GCMA Sakaridis et al. (2019)	81.7	46.9	58.8	22.0	20.0	41.2	**40.5**	**41.6**	64.8	31.0	32.1	**53.5**	47.5	**75.5**	39.2	0.0	49.6	30.7	21.0	42.0
MGCDA Sakaridis et al. (2020)	80.3	49.3	66.2	7.8	11.0	41.4	38.9	39.0	64.1	18.0	55.8	52.1	**53.5**	74.7	66.0	0.0	37.5	29.1	22.7	42.5
DANNet Wu et al. (2021)	90.4	60.1	71.0	33.6	**22.9**	30.6	34.3	33.7	**70.5**	**31.8**	80.2	45.7	41.6	67.4	16.8	0.0	**73.0**	31.6	22.9	45.2
CDAda Xu et al. (2021)	**90.5**	**60.6**	67.9	**37.0**	19.3	**42.9**	36.4	35.3	66.9	24.4	79.8	45.4	42.9	70.8	51.7	0.0	29.7	27.7	**26.2**	45.0
Ours	88.57	55.36	**74.44**	33.45	22.60	28.70	27.97	32.80	70.25	29.78	**81.40**	40.47	38.19	67.17	**72.32**	**4.79**	64.07	**36.89**	23.02	**46.96**

where N is the total number of valid pixels in the segmentation pseudo label, γ is the parameter of the focal loss, and p is the likelihood map of the correct classes. Unlike focal loss, we compute p at each pixel i in a 3×3 local region of class c:

$$p(c, i) = \max_{j} \left(o(c, j) \cdot P_{tn}^{S}(c, i) \right) \tag{8}$$

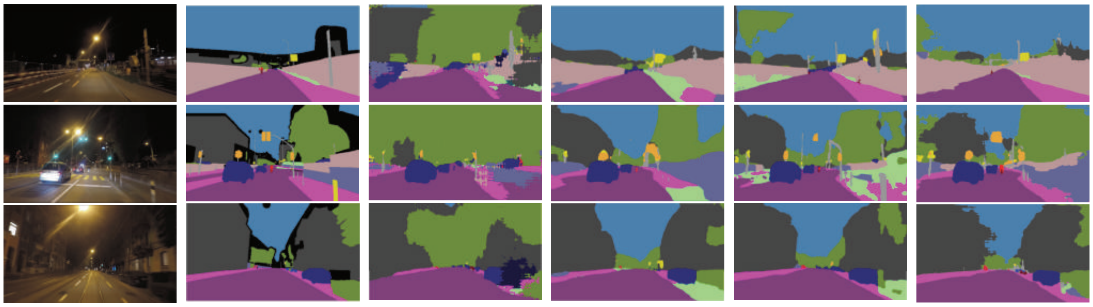
where o is the one-hot encoding of the pseudo-label $F_{td} = argmax(P_{td}^{S})$ and j represents each location of the 3×3 region centered at i.

Therefore, the total loss function is:

$$L_{total} = L_{seg} + L_{static} + \beta_1 L_{mix} + \beta_2 L_{kl} \tag{9}$$

where $\beta_1 = 0.04$, $\beta_2 = 0.75$. The values were decided by cross validation experiments.

4. Experiments



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

Figure 5: Visualization comparison of our framework with some existing state-of-the-art methods on three samples from Dark Zurich-val.

4.1. Datasets

Cityscapes The Cityscapes dataset Cordts et al. (2016) contains 5000 images captured in street with pixel-level annotations of 19 classes, and the resolution of both the original images and annotations is 2048 × 1024 pixels. There are 2975 images for training, 500 images for validation, and 1525 images for testing. In this paper, the Cityscapes dataset is used as the labeled source domain S.

Dark Zurich The Dark Zurich dataset Sakaridis et al. (2019) consists of 2416 night-time images, 2920 twilight images and 3041 daytime images, all of which are unlabeled and have a resolution of 1920 × 1080. Images can be coarsely aligned using a GPS-based nearest neighbor assignment algorithm. Only 2416 day-night image pairs are used in this paper to train the proposed night-time semantic segmentation framework. The Dark Zurich dataset also contains another 201 labeled night-time images, 50 for validation (Dark Zurich-val) and 151 for testing (Dark Zurich-test). Dark Zurich-test is an online benchmark whose ground truth are not public. In experiments, the performance of the proposed night-time segmentation framework in the Dark Zurich-test is obtained by submitting the segmentation results to an online evaluation website. This paper also uses Dark Zurich-val for method comparison and ablation study.

night-time Driving The night-time Driving dataset Dai and Gool (2018) contains 50 night-time images with a resolution of 1920 × 1080 from different scenes. All of these 50 images are annotated at the pixel level of 19 Cityscapes classes.

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

Following Wu et al. (2021), we train the model using the Stochastic Gradient Descent optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. The base learning rate is set to 2.5×10^{-4}, then a linear learning rate policy is used to reduce the learning rate by a factor of 0.9, and the batch size is set to 2. We apply a random crop of size 512 on the scale between 0.5 and 1.0 on the Cityscapes dataset, and a random crop of size 960 on the scale between 0.9 and 1.1 on the Dark Zurich dataset, and randomly horizontal flip.

The night-time semantic segmentation framework proposed in this paper is first compared with some existing state-of-the-art methods, including AdaptSeg Valada et al. (2017), ADVENT Vu et al. (2019), MGCDA Sakaridis et al. (2020), GCMA Sakaridis et al. (2019), DMAda Dai and Gool (2018), DANNet Wu et al. (2021) and CDAda Xu et al. (2021), and the results of mIoU performance are shown in Table I. The best results in the Table I are shown in bold, the second best results are underlined. Among these methods, the segmentation network of DANNet and our framework is PSPNet while others are ResNet-101. It can be seen that our method achieves the best performance among all methods, with an overall mIoU improvement of 1.7%. Our method has achieved improvements in the recognition of large objects such as the sky and buildings, and also achieved better results on small objects such as trucks, buses, trains and motorcycles, which shows that our framework can transfer more semantic knowledge of small size classes from day to night. To better illustrate the advantages of our approach, some visual examples are shown in Fig. 5.

The performance of the proposed method and other methods on the Night Driving-test is shown in Table II, and the example visualization results are shown in Fig. 6. Night Driving dataset is not as carefully labeled as the Dark Zurich-test. For different objects on this dataset, we get the following experimental results (mIoU): Road(88.42); Sidewalk(61.52); Building(88.54); Wall(39.27); Fence(0); Pole(47.08); Light(59.31); Sign(73.97); Vegetation(63.46); Terrain(0); Sky(54.15); Person(61.08); Rider(17.8); Car(60.49); Truck(38.99);

Bus(70.73); Train(80.06); Bicycle(34.15). As shown in Fig. 6, many categories for which our method predicts well, such as buildings, are not annotated in this test set. Even with these problems, our method still achieves the second best performance on this dataset (CDAda achieves the best performance).

Compared with DANnet, our model has higher stability in training process. As a GAN method, DANnet is difficult to train. Instead, the pretrained relighting network is used in our method in order to reduce the training complexity. When there is a large offset between between day-night image pairs, the effect of static loss in DANnet will be limited. We added the NightMix module to make better use of the correlation information in day-night image pairs. The higher performance of CDAda on Night Driving-test based on the fact that it uses more supervision information in the training process, including pre-classification and pseudo-labels. In contrast, our method does not rely on these additional supervision information. Therefore, our method has better adaptability for different datasets.

Table 2: Comparison with some state-of-the-art methods on night-time Driving-test.

Methods	mIoU
DMAda	36.1
GCMA	45.6
MGCMA	49.4
DANNet	47.7
CDAda	**50.9**
Ours	49.4

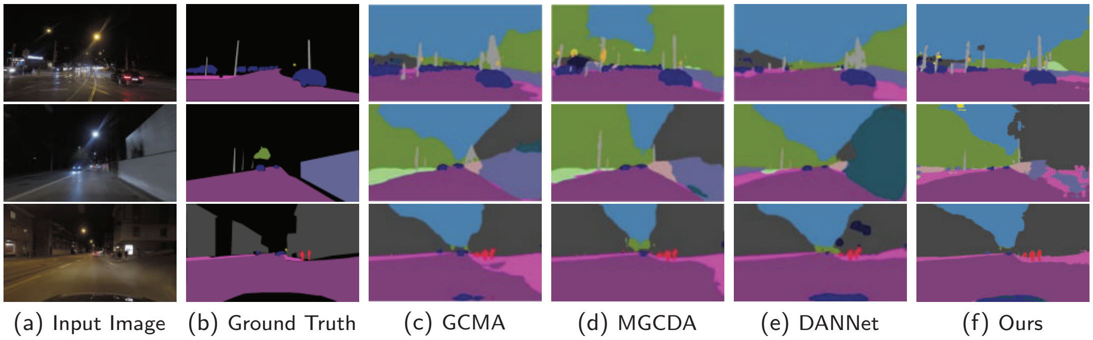


Figure 6: Visualization comparison of our framework with some existing state-of-the-art methods on three samples from Night Driving-test.

4.3. Ablation study

To demonstrate the effectiveness of the different modules of the proposed night-time semantic segmentation framework, several model variants are trained and tested on the Dark Zurich validation set, and the results of ablation experiments are shown in Table III. We find that the image relighting network can improve the performance, because the night-time

Table 3: Ablation study on several model variants of our framework on Dark Zurich-val.

Methods	mIoU
GCMA	26.65
MGCMA	26.10
DANNet	36.76
CDAda	36.00
Relight + L_{static}	34.19
R+L+NightMix	35.84
R+L+N+Consistency Reg	36.67
R+L+N+C+Cross Attention	**37.34**

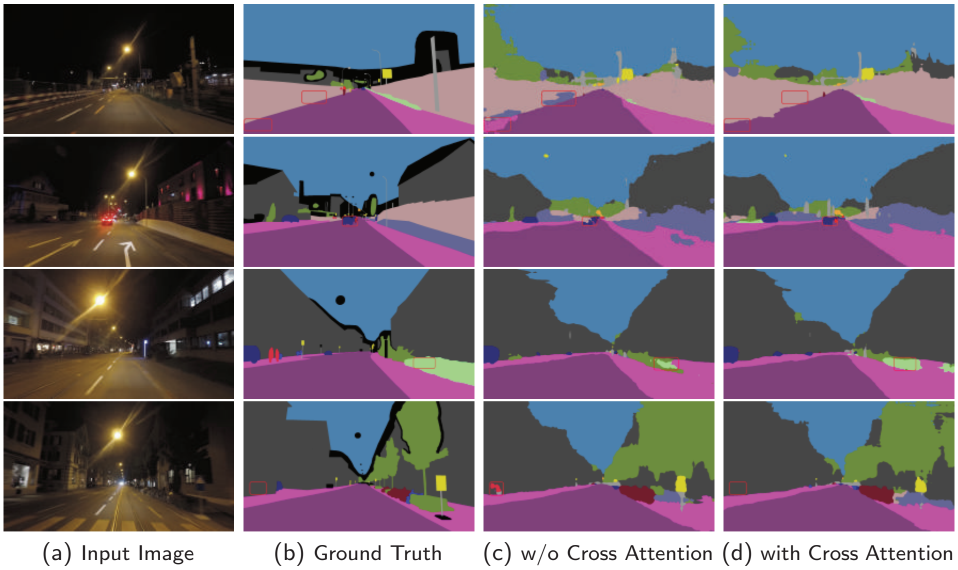


Figure 7: Visualization results of cross attention module on Dark Zurich-val. The cross attention module makes some details and object boundaries clear, such as the car in the second row. At the same time, some misclassified categories are now correctly classified, such as the walls in the first and fourth row.

images show more details after image relighting. The static loss L_{static} has a significant effect on performance improvement. The semantic information in the daytime scene can be transferred to the night-time scene through the static loss. Since the Dark Zurich dataset has no ground truth, the static loss can generate effective supervision information for night-time images. Although these supervision information is not completely correct, but it is enough to improve the performance of the model in night scenes. NightMix module has a great impact on the performance. Because there are no supervised semantic segmentation at night, the NightMix module can integrate supervised daytime scenes and unsupervised night-time scenes to improve the performance of night-time segmentation framework. Consistency regularization also improve the final performance of the framework, because it can

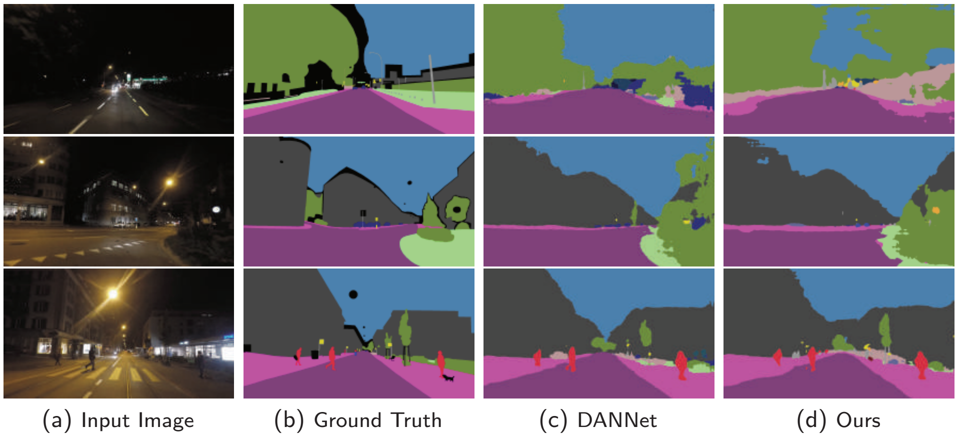


Figure 8: Visualization examples of poor performance in our framework on Dark Zurich-val. There are cluttered boundaries in our segmentation map.

resist the noise enhanced by image lighting. Cross attention module helps the model to establish associations between night-time features and daytime features, improving intra-class compactness, thereby improving the representations of semantic segmentation. The visualization results of cross attention module on Dark Zurich-val are shown in Fig. 7. The selective fusion of global features enhances the discrimination of details, the selective integration between day and night features helps to capture more contextual information, and the semantic consistency is significantly improved.

Although our framework has achieved good performance, it can be seen that there are cluttered boundaries in the segmentation map, as shown in Fig. 8. This is due to the poor visual conditions of the night scene, but also because the distribution of the segmentation map is not constrained in our framework. DANNet applies adaptive constraints to the distribution of the segmentation map through the discriminator, so the segmentation map generated by DANNet looks like the real segmentation map. Discriminator can also achieve a similar effect in our framework, but it will make the structure of the entire framework more complicated, and the training will be more troublesome and unstable. Perhaps the night-time semantic segmentation map can be constrained by the method of entropy constraint Vu et al. (2019) to make it closer to the real segmentation map.

5. Conclusions

In this paper, we propose a night-time semantic segmentation framework based on unsupervised learning and cross attention, which can utilize unlabeled night-time data and labeled daytime data to train a robust night-time semantic segmentation model. In order to perform night-time semantic segmentation without large-scale annotation datasets, we propose NightMix module, which integrates supervised daytime scenes and unsupervised night-time scenes. We use consistency regularization to enable the segmentation model to adapt to complex and changing night-time scene textures and lighting, while resisting additional noise enhanced by image relighting. We propose cross attention for the first time, which makes the model pay more attention to the parts of the night scene that are similar to the daytime scene, and establishes an association between the night-time features and the daytime features. Experimental results demonstrated the effectiveness of those modules and showed that our framework achieves the state-of-the-art performance on Dark-Zurich and Night Driving test datasets.

References

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille, editors. *Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs*, volume abs/1412.7062 of *CoRR*, 2015.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, editors. *Rethinking Atrous Convolution for Semantic Image Segmentation*, volume abs/1706.05587 of *ArXiv*, 2017a.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille, editors. *DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs*, volume 40 of *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Y. Wang, and Min Sun, editors. *No More Discrimination: Cross City Adaptation of Road Scene Segmenters*, 2017 IEEE International Conference on Computer Vision (ICCV), 2017b.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, editors. *The Cityscapes Dataset for Semantic Urban Scene Understanding*, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

Dengxin Dai and Luc Van Gool, editors. *Dark Model Adaptation: Semantic Image Segmentation from Daytime to Nighttime*, 2018 21st International Conference on Intelligent Transportation Systems (ITSC), 2018.

J. Fu, J. Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu, editors. *Dual Attention Network for Scene Segmentation*, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Tak Wu Kwong, and Runmin Cong, editors. *Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement*, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu, editors. *Squeeze-and-Excitation Networks*, volume 42 of *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger, editors. *Computer Vision for Autonomous Vehicles: Problems, Datasets and State-of-the-Art*, volume 12 of *Found. Trends Comput. Graph. Vis.*, 2020.

Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross B. Girshick, editors. *PointRend: Image Segmentation As Rendering*, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár, editors. *Focal Loss for Dense Object Detection*, volume 42 of *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox, editors. *Semi-Supervised Semantic Segmentation With High- and Low-Level Consistency*, volume 43 of *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson, editors. *ClassMix: Segmentation-Based Data Augmentation for Semi-Supervised Learning*, 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021.

Brian Paden, Michal Cáp, Sze Zheng Yong, Dmitry S. Yershov, and Emilio Frazzoli, editors. *A Survey of Motion Planning and Control Techniques for Self-Driving Urban Vehicles*, volume 1 of *IEEE Transactions on Intelligent Vehicles*, 2016.

Joseph Redmon and Ali Farhadi, editors. *YOLOv3: An Incremental Improvement*, volume abs/1804.02767 of *ArXiv*, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox, editors. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, 2015.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool, editors. *Guided Curriculum Model Adaptation and Uncertainty-Aware Evaluation for Semantic Nighttime Image Segmentation*, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool, editors. *Map-Guided Curriculum Domain Adaptation and Uncertainty-Aware Evaluation for Semantic Nighttime Image Segmentation*, volume PP of *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Wilko Schwarting, Javier Alonso-Mora, and Daniela Rus, editors. *Planning and Decision-Making for Autonomous Vehicles*, 2018.

Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun, editors. *OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks*, volume abs/1312.6229 of *CoRR*, 2014.

Evan Shelhamer, Jonathan Long, and Trevor Darrell, editors. *Fully Convolutional Networks for Semantic Segmentation*, volume 39 of *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin Dogus Cubuk, Alexey Kurakin, Han Zhang, and Colin Raffel, editors. *FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence*, volume abs/2001.07685 of *ArXiv*, 2020.

Nasim Souly, Concetto Spampinato, and Mubarak Shah, editors. *Semi Supervised Semantic Segmentation Using Generative Adversarial Network*, 2017 IEEE International Conference on Computer Vision (ICCV), 2017.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, editors. *Deep High-Resolution Representation Learning for Human Pose Estimation*, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

Abhinav Valada, Johan Vertens, Ankit Dhall, and Wolfram Burgard, editors. *AdapNet: Adaptive semantic segmentation in adverse environmental conditions*, 2017 IEEE International Conference on Robotics and Automation (ICRA), 2017.

Tuan-Hung Vu, Himalaya Jain, Max Bucher, Matthieu Cord, and Patrick Pérez, editors. *ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation*, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

Xinyi Wu, Zhenyao Wu, Haojie Guo, Lili Ju, and Song Wang, editors. *DANNet: A One-Stage Domain Adaptation Network for Unsupervised Nighttime Semantic Segmentation*, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

Zifeng Wu, Chunhua Shen, and Anton van den Hengel, editors. *Wider or Deeper: Revisiting the ResNet Model for Visual Recognition*, volume abs/1611.10080 of *ArXiv*, 2019.

Qi Xu, Yinan Ma, Jing Wu, Chengnian Long, and Xiaolin Huang, editors. *CDAda: A Curriculum Domain Adaptation for Nighttime Semantic Segmentation*, 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021.

Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang, editors. *SegFix: Model-Agnostic Boundary Refinement for Segmentation*, volume abs/2007.04269 of *ArXiv*, 2020.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, editors. *Pyramid Scene Parsing Network*, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba, editors. *Semantic Understanding of Scenes Through the ADE20K Dataset*, volume 127 of *International Journal of Computer Vision*, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, editors. *Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks*, 2017 IEEE International Conference on Computer Vision (ICCV), 2017.