applied sciences

MDPI

Article

An Adaptive Semantic Segmentation Network for Adversarial Learning Domain Based on Low-Light Enhancement and Decoupled Generation

Meng Wang [1,*,†], Zhuoran Zhang [1,†] and Haipeng Liu [2,†]

[1] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China; zzr@stu.kust.edu.cn
[2] Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, China; ran@kust.edu.cn
* Correspondence: wangmeng@kmust.edu.cn
† These authors contributed equally to this work.

Abstract: Nighttime semantic segmentation due to issues such as low contrast, fuzzy imaging, and low-quality annotation results in significant degradation of masks. In this paper, we introduce a domain adaptive approach for nighttime semantic segmentation that overcomes the reliance on low-light image annotations to transfer the source domain model to the target domain. On the front end, a low-light image enhancement sub-network combining lightweight deep learning with mapping curve iteration is adopted to enhance nighttime foreground contrast. In the segmentation network, the body generation and edge preservation branches are implemented to generate consistent representations within the same semantic region. Additionally, a pixel weighting strategy is embedded to increase the prediction accuracy for small targets. During the training, a discriminator is implemented to distinguish features between the source and target domains, thereby guiding the segmentation network for adversarial transfer learning. The proposed approach's effectiveness is verified through testing on Dark Zurich, Nighttime Driving, and CityScapes, including evaluations of mIoU, PSNR, and SSIM. They confirm that our approach surpasses existing baselines in segmentation scenarios.

Keywords: domain adaptation; nighttime semantic segmentation; adversarial learning; low-light enhancement

Citation: Wang, M.; Zhang, Z.; Liu, H. An Adaptive Semantic Segmentation Network for Adversarial Learning Domain Based on Low-Light Enhancement and Decoupled Generation. *Appl. Sci.* **2024**, *14*, 3295. https://doi.org/10.3390/app14083295

Academic Editor: Tobias Meisen

Received: 2 March 2024
Revised: 28 March 2024
Accepted: 9 April 2024
Published: 13 April 2024

1. Introduction

Semantic segmentation is a fundamental task in computer vision where each pixel of a given image is labeled with an object category. It is widely used in various applications such as autonomous driving [2], medical imaging [3], and human parsing [4]. In recent years, the performance of semantic segmentation of daytime scene images has substantially improved due to the rapid progress in deep learning and computing power. As researchers have tackled more challenging image segmentation scenarios under various limited, adverse, and degraded conditions, semantic segmentation of nighttime images [5] has emerged as a prominent research focus. However, nighttime semantic segmentation poses unique challenges; for example, low contrast of the input images makes it difficult to obtain clear and complete segmentation boundaries, and the variation in lighting conditions might lead to changes in the brightness and color of the objects within the same scene. Additionally, the manual labeling of a high-quality training set of nighttime images is also a formidable task, contributing to the degradation of segmentation model performance. The present study seeks to address the above-mentioned bottlenecks via proposing a nighttime semantic segmentation network that is suitable for real-scenario applications such as autonomous driving and security monitoring.

Nighttime images, as a class of low-light images, have many regions of foreground pixels that are not obvious or recognizable to the human eye, and it is difficult to perform

high-quality pixel-level annotation on this part of the image. Consequently, a sufficient quantity of accurate segmentation instances is the basis for realizing efficient learning of segmentation models. To tackle this, current schemes such as domain adaptation [6,7], synthetic datasets [8], and style transfer [9] are commonly employed. Due to the low brightness and contrast of nighttime images, this paper transfers daytime images to nighttime images by domain adaptation. However, large differences in scene feature distributions and foreground types between the source and target domains, primarily concerning light intensity, can often lead to the distortion of crucial spatial semantic details during domain transfer. In view of this, some studies have proposed to establish domain transfer from the daytime domain to the nighttime domain using an intermediate domain as a smooth transition, such as the twilight domain. In [5,10,11], the twilight domain serves as a bridge, allowing the model trained in the daytime domain to adapt progressively to the nighttime domain by extracting features from twilight images and performing transfer alignment learning. In [12], a model adaptation method based on course learning is proposed, which adapts the model to light changes and noise in nighttime scenes by gradually increasing the complexity of the nighttime images. Building [12], research conducted in [13] utilizes feature maps to provide prior knowledge about nighttime scenes, aiding the model in understanding objects and structures and guiding adaptive training. In [14], an encoder–decoder structure for semantic segmentation of nighttime images is introduced which uses a domain map approach for mapping synthetic to real data. In [15–17], a generator network is trained using adversarial learning to translate daytime images to nighttime images. Subsequently, the feature extractor of the generator network is adopted for the semantic segmentation network to extract transform-based regularized features from nighttime images. Furthermore, to enhance the generalization performance of the model, research in [18] employs Adversarial Generative Networks (GANs) to translate daytime images to nighttime images, and random transformations are then applied to those images, followed by joint training using the adversarial and semantic segmentation loss functions.

Among the above-mentioned methods, refs. [5,10–13] leverage intermediate domains to create a smooth transition, thereby improving model generalization and potentially reducing dataset labeling costs. Nonetheless, introducing intermediate domains may entail additional preprocessing and model training and fail to fully cover all variations from daytime to nighttime. On the other hand, refs. [14–16,18] employ techniques such as style transfer and build synthetic datasets to address the difficulty of labeling nighttime (low-light) samples. While building synthetic datasets offers advantages, it also carries the risk of introducing bias and noise. Furthermore, these schemes focus only on the statistical representation of the overall image style in style transfer and thus are prone to the loss of spatial details. In addition, generating a transferred image with the same semantics as the original image, especially when dealing with a relatively large domain gap, remains a challenging aspect of image translation.

To address these problems, this paper uses pairs of day and night images in similar scenes as target domains and tries to transfer the source domain generalized model to the scene-specific multi-target domain without introducing an intermediate domain, synthetic datasets, or style migrations so as to improve the segmentation quality by joint adversarial learning and multi-domain co-training. Based on this, this paper proposes an adaptive semantic segmentation network for the adversarial learning domain based on low-light enhancement and decoupled generation (DLA-Net). At the front end of the model, a lightweight low-light image enhancement network (LIE-SubNet) is embedded to elevate foreground contrast in nighttime images and accomplish spatial feature alignment across different illuminance datasets. Existing segmentation models typically treat the foreground target as a unified entity; however, foreground boundary regions usually contain richer spatial details with higher-frequency feature information, whereas non-boundary regions exhibit fewer spatial details featured by low-frequency distributions. Inspired by [6], this paper leverages a generative network capable of decoupling the foreground body and edge to predict segmentation masks and uses two discriminators for adversarial training between

the source and target domains. Additionally, a small pixel reweighting strategy [19] is implemented to process the input images and reduce prediction uncertainty, thus improving the segmentation accuracy for small targets. In our experiments, the Dark Zurich dataset [5] is employed, which contains pairs of daytime and nighttime images based on rough GPS positional alignment. Through extensive testing on the Dark Zurich, CityScapes, and Nighttime Driving [10] datasets, the proposed method is verified to demonstrate improved performance in low-light nighttime semantic segmentation. The primary contributions of our work are summarized as follows:

- In this paper, a multi-domain model joint training network for semantic segmentation, DLA-Net, is introduced which transfers the source domain to the multi-target domain of a specific scene without requiring an intermediate domain. It accomplishes joint adversarial training of the multi-domain model, supported by the low-light image enhancement sub-network, on the multi-target domain;
- The low-light image enhancement sub-network, LIE-SubNet, which combines deep learning and mapping curve iteration, is proposed to enhance pixel contrast and spatial feature alignment of nighttime images. In the segmentation network, a generative network capable of decoupling subjects and edges is utilized to guide segmentation prediction via exploiting the adversarial loss in the daytime and nighttime domains;
- To effectively utilize both low-frequency and high-frequency information of foreground targets, the segmentation mask is decoupled into the body generation branch and the edge preservation branch. These branches can focus on different attributes of the regional features during training. The resulting masks are then composited and reconstructed to achieve a complete semantic segmentation mask capable of retaining the details while removing the void noise.

2. Related Work

Domain adaptation for semantic segmentation: Domain adaptation seeks to transfer knowledge learned in the source domain to the target domain, where the object classes are similar but the distribution of data statistics differs. Currently, a portion of domain adaptive schemes adopt adversarial learning frameworks, introducing an adversarial loss function between the source and target domains to guide the model in aligning feature representations across different data domains. For example, in [20], Hoffman et al. proposed a new approach to semantic segmentation using category-constrained [21] full convolutional domain adversarial learning. AdaptSegNet [6] utilizes adversarial training to achieve feature alignment in the source and target domains. Additionally, several approaches employ joint training and multiple task learning strategies to improve model generalization by sharing parameters among source and target domains. In BDL [22], images from both source and target domains are input into a shared convolutional neural network, with the last layer divided into two branches for semantic segmentation tasks in the respective source and target domains. Through sharing the feature extraction layer of the network, the source and target domains can leverage the underlying image feature representation, thereby improving adaptation to redundant representations of the target domain.

Unlike adversarial learning, style transfer [18] and image translation from source image to target image are also widely used for domain adaptation. They typically incorporate a domain invariance loss function into the generator network to enforce domain-invariant image generation in the target domain [23,24]. This type of loss function commonly comprises both an adversarial loss and a domain invariance loss. Specifically, the generator network is trained using the adversarial loss, while the domain invariance loss is used to teach the generator network to learn the shared features between source and target domains, resulting in domain-invariant representations. Some other studies have explored the combination of self-training strategy and fine-tuning strategy through multiple rounds of network training. However, the self-training strategy may introduce noise when using pseudo-labeling, thus impacting model performance [25]. To mitigate the influence of

noise, researchers have proposed several improved self-training techniques [26,27], such as using model ensembles to reduce noise or refining pseudo-labeling to reduce mislabeling. Alternatively, some studies have employed course-based learning [28,29] to acquire simple attributes in the target domain before using them to normalize semantic segmentation models. However, the significant visual disparities between daytime and nighttime images pose a formidable challenge for these methods, making them ill suited to effectively handling domain adaptation in scenarios with markedly different illumination intensities. Consequently, they often fall short of delivering satisfactory performance in nighttime semantic segmentation. This paper explores more efficient techniques to minimize the domain gap so that transfer models can achieve accurate segmentation predictions.

Nighttime (low-light) semantic segmentation: Some studies have demonstrated the effectiveness of employing intermediate domains for the progressive adaptation of semantic models trained on daytime scenes to nighttime scenes. For example, Dai et al. [10] proposed a step-by-step adaptive approach based on intermediate domains. This approach leverages an intermediate twilight domain as a bridge between daytime and nighttime scenes and trains an intermediate model on the twilight domain, which is then applied to the semantic segmentation of nighttime scenes. Later, Sakaridis et al. [5,6] extended the approach to a class of guided curriculum adaptation frameworks, incorporating synthetic and unlabeled real images to establish correspondences in scene images at various times. However, it is worth noting that this progressive adaptation approach often necessitates training several semantic segmentation models. For instance, in [5], three models were trained separately for three different domains, potentially making the training process less efficient. Building upon this methodology of using intermediate domains for progressive domain adaptation, some studies have trained some additional image translation models. CycleGAN [18] is a good case in point and enables the inter-transfer of daytime and nighttime images before training semantic segmentation models, thus introducing different visual features through diverse augmented data and aiding in the adaptation of the transferred models to various scenes and environments. Furthermore, ref. [30] introduced a nighttime semantic segmentation method based on image translation by translating nighttime images into daytime ones and utilizing semantics with the model trained on the daytime domain.

More recently, to improve the semantic segmentation of night scenes, researchers have explored the use of different sensors to capture the same image as an auxiliary input. Vertens et al. [31] proposed to utilize the insensitivity of thermal infrared to changes in illumination as a supplemental input to segmented images to provide additional information for nighttime semantic segmentation. Additionally, other studies have devised specialized scene semantic segmentation methods. For example, Ref. [32] proposed a two-stage adversarial training approach that employs domain adaptation techniques to transform between pairs of daytime and nighttime scenes, particularly for rainy and nighttime scenarios. Likewise, Ref. [33] introduced an adaptive network capable of automatically adapting its internal architecture based on the attributes of input images to different environmental conditions, including nighttime and rainy ones. Differing from the above methods, this paper introduces a network structure designed to train semantic segmentation for low-light images via end-to-end adversarial learning without resorting to intermediate domains or auxiliary images.

3. Method

3.1. Framework Overview

The domain adaptive method proposed in this paper involves two key domains: a source domain S for pre-training, which can be any normal lighting scene, and a target domain $T = \{T_d, T_n\}$ containing two roughly aligned subdomains, T_d and T_n, representing daytime and nighttime scenes, respectively. In the pre-training phase, a labeled image set $S \triangleq \{X_S, Y_S\}$ from the source domain was used to optimize the semantic segmentation network parameters. Subsequently, two discriminators, $D_{S \to T_d}$ and $D_{S \to T_n}$, were employed to bootstrap the domain adaptive model transfer from S to T_d and from S to T_n to efficiently

model semantic segmentation of the nighttime scene T_n in the target domain. The domain adaptive semantic segmentation network in this paper comprises three modules: (1) a low-light image enhancement network N_{en}, (2) a pre-trained semantic segmentation network G_{pre} and a transfered semantic segmentation network G_{seg}, which decouples the body and edge during segmentation and provides predicted image dimensions of $\mathbb{R}^{H\times W\times C}$, with C denoting the total number of image categories, and (3) a segmented mask activation network N_Y, which consists of a convolutional layer and a sigmoid normalization function, as shown in Figure 1. The network input contains the source domain image X_S and the target domain images X_{T_d} and X_{T_n}, consisting of three types of domain samples. Among them, X_{T_n} was additionally passed through a nighttime (low-light) enhancement network N_{en}, which generated an enhancement loss L_{en} to optimize the enhancement result and brought the output closer to the daytime domain. The network uses image annotations X_S, Y_S in the source domain S dataset to compute the segmentation loss L_{seg} and then obtains the segmentation prediction masks $\tilde{F} = \{\tilde{F}_S, \tilde{F}_{T_d}, \tilde{F}_{T_n}\}$ and segmentation loss L_{de} by G_{seg}. After that, two discriminators, $D_{S\to T_d}$ and $D_{S\to T_n}$, perform adversarial transfer learning, and the final segmentation masks $\hat{Y} = \{\hat{Y}_S, \hat{Y}_{T_d}, \hat{Y}_{T_n}\}$ are obtained via activating the network N_Y, i.e., $\hat{Y} = \{N_Y(\tilde{F}_S), N_Y(\tilde{F}_{T_d}), N_Y(\tilde{F}_{T_n})\}$. The whole network guides the domain adaptive alignment of the model based on the composite total loss L_{total}.

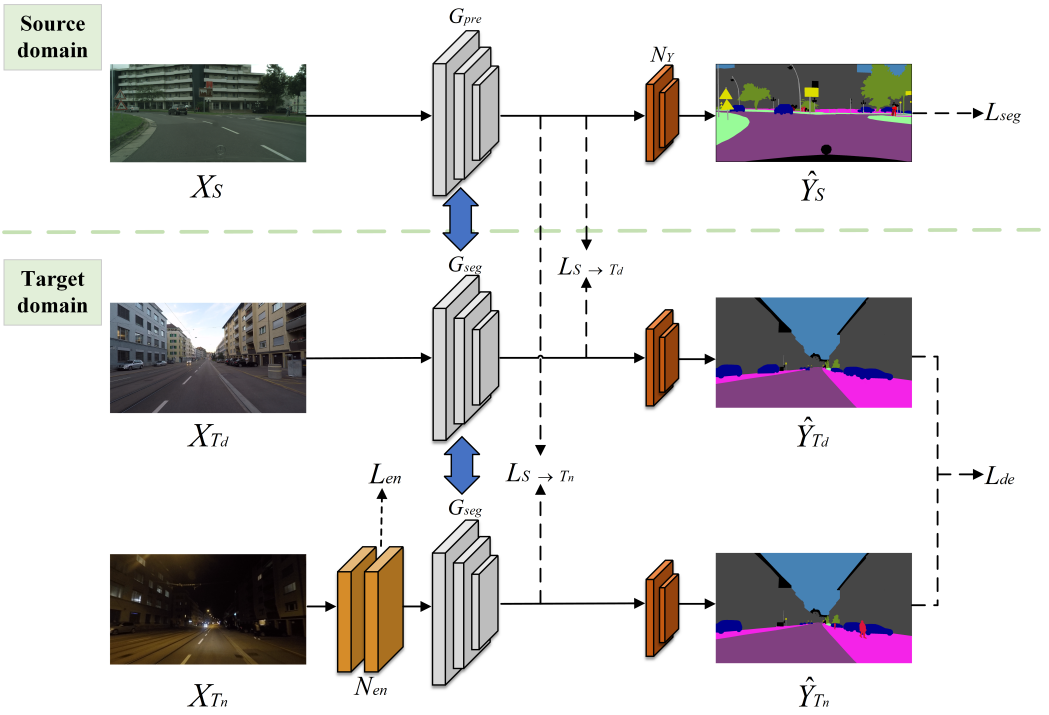


Figure 1. Overall structure of the network proposed in this paper (DLA-Net). The network takes three types of domain-related samples as input: source domain image X_S and target domain images X_{T_d} and X_{T_n}. Within the framework, $L_{S\to T_d}$ and $L_{S\to T_n}$ are the adversarial losses of S and T_d, while S and T_n are obtained from $D_{S\to T_d}$ and $D_{S\to T_n}$, respectively.

3.2. Low-Light Image Enhancement Sub-Network

In the realm of image illumination enhancement, the majority of research commonly employs methods like mapping curves or neural networks. However, this paper has the initiative to fit mapping curves with neural networks to design a low-light image enhancement sub-network. The objective was to homogenize the intensity distribution of the input image X_{T_n} from the nighttime target domain T_n and generate the enhanced image $\hat{X}_{T_n}$, ensuring that the predictions of different domain samples align after passing through the segmentation network. Inspired by [14], we utilized an iterative pixel enhancement map-

ping curve to adjust the brightness and contrast of the image through the pixel grayscale mapping relationship, as shown in Equation (1).

$$Ie(X_{T_n}(x);\alpha) = log(1 + X_{T_n}(x) + \alpha X_{T_n}(x)(1 - X_{T_n}(x))), \quad (1)$$

where x is the pixel coordinate, and the α parameter ensures that each pixel value in the enhanced image falls within the normalized range of $[0,1]$, preventing any loss of information due to overflow. By setting α to a value between -1 and 1, the Ie curve can be controlled within the range of $[0,1]$. For example, when $\alpha = -1$, $Ie(X_{T_n}(x);-1) = log(1 + X_{T_n}(x)^2)$, i.e., each value is within $[0,1]$.

To adapt to more challenging low-light conditions, iterating the quadratic curve Ie could result in a higher-order curve. Although the higher-order curve is able to adjust the image over a wider area, it still applies a global adjustment as the α value is applied to all pixels, resulting in over-enhancement or diminution of localized regions. To solve this problem, we used a separate curve for each RGB channel of the input image to perform an iterative transformation so that each channel has a corresponding optimal α value for image enhancement, as shown in Equation (2).

$$Ie_m(X_{T_n}(x);\mathcal{A}) = log(1 + Ie_{m-1}(X_{T_n}(x)) + \mathcal{A}_m Ie_{m-1}(X_{T_n}(x))(1 - Ie_{m-1}(X_{T_n}(x)))), \quad (2)$$

where m, set to 8 in this paper, signifies the number of iterations and controls the curvature, and $\mathcal{A}$ is a parametric mapping with the same size as the given image used to represent the optimal α value for each channel. To obtain the mapping relationship among the input image and its optimal curve parameter mapping, this paper proposes a depth curve fitting network, as illustrated in Figure 2.

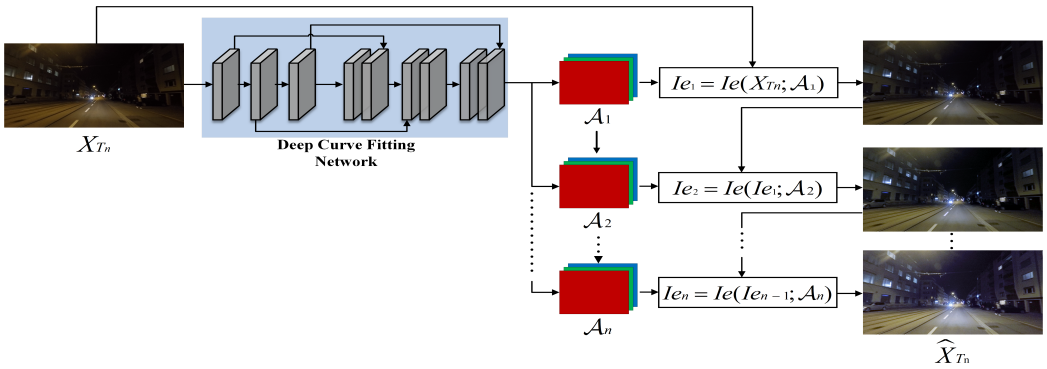


Figure 2. Architecture of the LIE-SubNet architecture. The network was designed to evaluate a set of optimal light enhancement curves (Ie curves) that iteratively enhance the input image. The deep curve fitting network uses an ordinary CNN with six alternately connected convolutional layers, each consisting of 32 3×3 convolutional kernels with a step size of 1. A ReLu function is added at the end of the network.

To evaluate the quality of the enhancement image, we used the following three losses to train the image enhancement network.

To suppress overexposure or underexposure of certain areas, we designed an exposure control loss L_{ec} to regulate the level of exposure. L_{ec} quantifies the disparity between the mean luminance value of a specific area and the intended exposure level e. e was set to a grayscale value in the RGB color space following existing methods [34,35] in this paper. This loss brought the enhancement closer to the desired exposure level, mitigated overexposure or underexposure, and hence obtained a more visualized and higher-quality image, as shown in Equation (3).

$$L_{ec} = \frac{1}{V}\sum_{i=1}^{V}|\hat{I}_i - e|, \quad (3)$$

where V denotes the number of non-overlapping regions with a size of 16×16, and $\hat{I}$ represents the average luminance value of localized region V in the augmented image $\hat{X}_{T_n}$. e was set to 0.5 in the experiment.

The color constancy loss employed in this paper was based on the Gray-World [36] color constancy assumption, which posits that each color channel is averaged as gray over the whole image. This loss rectifies potential color deviations in the enhanced image, recovers color information affected by changes in illumination, improves the quality and visual perception of the image, and determines the relationship among the three color channels, as shown in Equation (4).

$$L_{cc} = \sum_{\forall (a,b) \in \tau} (\bar{E}^{(a)} - \bar{E}^{(b)})^2, \tag{4}$$

where $\bar{E}^{(a)}$ and $\bar{E}^{(b)}$ denote the average intensity values of the a-channel and the b-channel, respectively, in the enhanced image $\hat{X}_{T_n}$, and (a, b) denotes a pair of channels, $\tau = \{(R, G), (R, B), (G, B)\}$. The smaller value of L_{cc} indicates that the color of the brightened image is more balanced, and the larger L_{cc} indicates that the brightened image may have the problem of color bias.

In this paper, an illumination smoothness loss [37] was built into each curve parameter mapping $\mathcal{A}$ to maintain a monotonic relationship between adjacent pixels. The loss assists the model in learning that the illumination changes in the neighboring regions exhibit both consistency and smooth transition and improving image processing performance and image quality. It is shown in Equation (5).

$$L_{is} = \frac{1}{M} \sum_{m=1}^{M} \sum_{s=1}^{H \times W \times C} \left(\left| \nabla_x \mathcal{A}_m^{(s)} \right| + \left| \nabla_y \mathcal{A}_m^{(s)} \right| \right)^2, \tag{5}$$

where M stands for the number of iterations. Specifically, C denotes the RGB color channels, and $C = 3$. ∇_x and ∇_y denote the horizontal and vertical gradient operations, respectively. The smaller the value of L_{is}, the smoother the light of the brightened image, and vice versa, which indicates that there are mutations or artifacts in the light of the brightened image.

The total enhancement loss is shown in Equation (6).

$$L_{en} = L_{ec} + \lambda_1 L_{cc} + \lambda_2 L_{is}, \tag{6}$$

where λ_1 and λ_2 are hyperparameters used to balance the size of the loss and were set to 0.5 and 20 in the experiments, respectively.

Contributing to the realm of LIE-SubNet, this paper explores the combination of a set of higher-order curves that can be iterated with a deep learning network for different numbers of iterations to verify the optimal performance and enhance nighttime pixel contrast. The method reduces the domain gap among the daytime and nighttime domains without resorting to an intermediate domain or the training of multiple distinct models and feeds the segmentation network with smaller differences in illumination images.

3.3. Semantic Segmentation Network for Decoupling Body and Edge

Currently, mainstream semantic segmentation methods primarily focus on enhancing the internal consistency of the object through global modeling or refining the object details along the boundaries through multi-scale feature fusion. However, it is worth noting that foreground boundary regions typically harbor more spatial detail and higher-frequency feature information. In view of this, we introduced the semantic segmentation network G_{seg} for decoupling body and edge, which contains a body generation branch ρ and an edge preservation branch δ. Unlike previous studies, we do not require the input image's ground truth map and trained two branches with distinct losses to predict the body feature map and edge feature map, respectively. The implementation details are described below.

Decoupling segmentation framework: In this paper, we assume that the spatial features of the image conform to the addition rule, i.e., $\tilde{F} = F_{\text{body}} + F_{\text{edge}}$. Accordingly, the body feature F_{body} can be generated first, and the edge feature F_{edge} can be obtained by a specific subtraction operation. If we make $F_{body} = \rho(F)$, then $F_{edge} = F - F_{body}$, as shown in Equation (7).

$$\begin{aligned} \tilde{F} &= \rho(F) + \delta(F_{edge}) \\ &= F_{body} + \delta(F - F_{body}), \end{aligned} \tag{7}$$

where ρ represents the body generation branch mapping which is used to aggregate contextual information within objects to form a distinct body for each object. On the other hand, δ denotes the edge preservation branch mapping, which is designed to extract spatially detailed features from the boundary region.

Body generation branch: This branch is responsible for the generation of more consistent feature representations for pixels that are part of the same object in an image. Low-resolution feature maps typically contain low-frequency terms, with the low-spatial-frequency portion representing the image as a whole. Therefore, the low-resolution feature maps represent the most salient parts. In order to achieve this goal, as illustrated in Figure 3, X is the input image, and we utilized an encoder–decoder architecture after the backbone to extract F. Specifically, the encoder downsamples F using dilated convolution, which downsamples F into a low-resolution representation of the low-spatial-frequency portion, denoted as F_{low}. In some cases, low-resolution features might still contain high-frequency information. We assumed that this compressed representation encapsulates the most obvious object portions and leads to rough representation which ignores details or high-frequency portions. Therefore, we used bilinear interpolation to upsample F_{low} to the same size as F to obtain F_{up}. Then, we cascaded F and F_{up} and used a $1 \times 1Conv$ to adjust the channel dimensions to $\mathbb{R}^{H \times W \times C}$ to obtain F_{conv}, i.e., $F_{conv} = h_{conv}(F||F_{up})$, where h_{conv} denotes the 1×1 convolutional layer, and $||$ denotes the channel dimensionality join operation. This branch also contains an average pooling layer by average pooling F_{conv} to generate a feature map F_{ap} with a more distinct body, i.e., $F_{ap} = h_{ap}(F_{conv})$ and $F_{ap} \in \mathbb{R}^{H \times W \times C}$, where h_{ap} denotes the average pooling operation.

To increase the spatial accuracy of body features in segmentation results, we first mapped each pixel p in the default spatial grid Ω_l on F_{ap} to a new pixel point p via feature relocation. Then, we used a variable bilinear sampling mechanism [38,39] to approximate the value of each pixel point p in F_{body}, i.e., $F_{body}(\hat{p}) = \sum_{p \in l(o)} F_{ap}(p)$, where l denotes the pixels in the four fields around p, o is the center point, and $F_{body} \in \mathbb{R}^{H \times W \times C}$. In addition, to ensure smoother performance of the body feature and reduce noise and discontinuities in the prediction results, we applied the L_2 loss [40] to bootstrapping the body generation branch learning, as shown in Equation (8).

$$L_{body} = \sqrt{\sum_{s=1}^{H \times W \times C} (F_{body}^{(s)})^2}, \tag{8}$$

where s denotes the positional index of the element in $\forall F \in \mathbb{R}^{H \times W \times C}$, and $s = 1, 2, ..., H \times W \times C$.

Edge preservation branch: This branch is dedicated to handling high-frequency terms F_{high} in the image, where high-frequency features usually encompass more detailed edge information. To obtain the high-frequency edge feature, we subtracted the body feature F_{body} from the original feature F, i.e., $(F - F_{body})$. Drawing inspiration from recent work on decoder design [41], we outputted a low-level feature F_{detail} through the backbone's low layer, which served as a complement to the missing fine-detail information and augmented the high-frequency terms in F_{edge}. Finally, $(F - F_{body})$ and F_{detail} were cascaded, and then a $1 \times 1Conv$ was used for channel adjustment to obtain F_{edge}. The implementation is expressed in Equation (9).

$$F_{edge} = h_{\text{conv}}((F - F_{body})||F_{detail}), \tag{9}$$

where $F_{edge} \in \mathbb{R}^{H \times W \times C}$.

The edge preservation branch focuses more on edge detail features and does not require body features. Unlike the L_2 loss, the L_1 loss can obtain a sparse solution so that certain features have zero weight. This makes the boundary sparser and reduces unnecessary body features, contributing to an accurate boundary prediction feature map. Therefore, the L_1 loss [40] was utilized to guide the learning of the edge preservation branch, as shown in Equation (10).

$$L_{edge} = \sum_{s=1}^{H \times W \times C} |F_{edge}^{(s)}|. \tag{10}$$

The final decoupling loss is:

$$L_{de} = L_{body} + \lambda_3 L_{edge}, \tag{11}$$

Both the L_{body} and L_{edge} losses complement each other by sampling pixels from different regions of the image, which was beneficial for showing the performance of the experimental results. Since the edge portion is not a large part of the overall image, λ_3 is used to balance the weight of L_{edge} in L_{de}, which was set to 0.4 in the experiments.

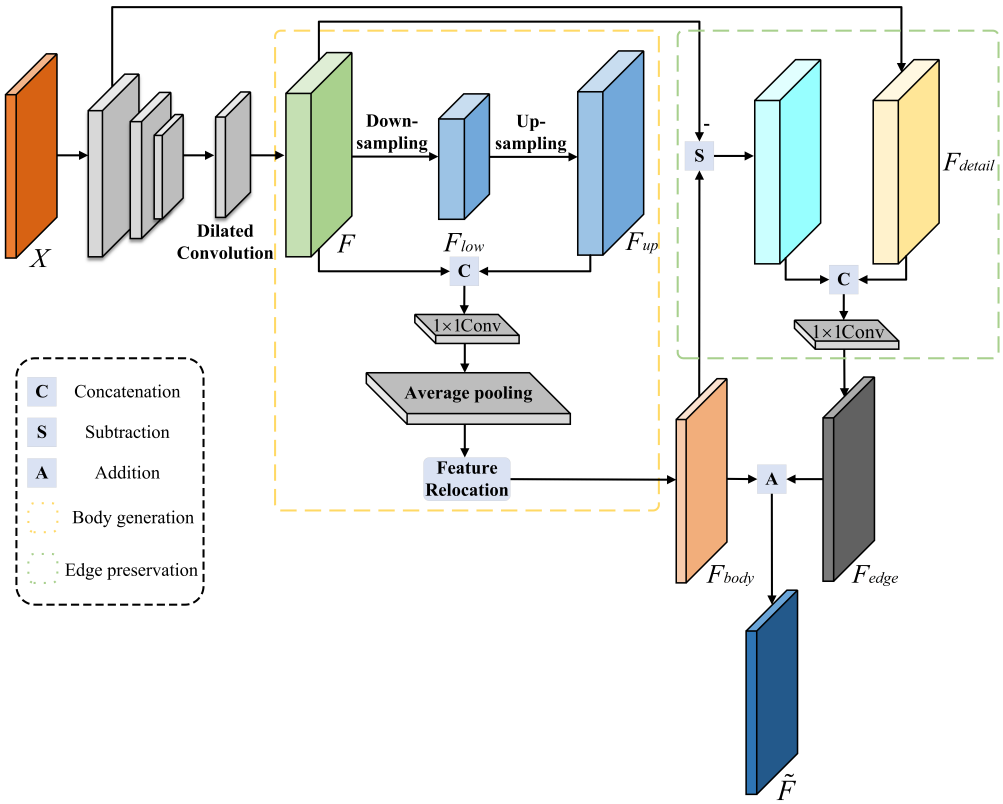


Figure 3. Decoupling module for body generation and edge preservation. In this module, X represents the input image, F is derived from the backbone network and a dilated convolution, and F_{detail} represents the high-frequency detailed feature which is output through the low layer of the backbone network. In the body generation branch, F and F_{up} are cascaded and input to $1 \times 1Conv$. Notably, F_{low} was not added to the cascade. Subsequently, average pooling and feature relocation were performed to obtain the body feature F_{body}, which had an obvious body but fuzzy edges. In the edge preservation branch, $(F - F_{body})$ was cascaded with F_{detail} and input to $1 \times 1Conv$ to obtain the edge feature F_{edge} with a clear boundary. The final segmentation predicted $\tilde{F} = F_{body} + F_{edge}$.

In this paper, G_{seg} acquired the body feature map F_{body} and edge feature map F_{edge} of the input image X through the body generation branch ρ and the edge preservation branch δ, respectively. Moreover, the edge features were supplemented by the high-frequency detailed features F_{detail} output from the lower layer of the backbone network. By employing distinct body and edge losses, the segmentation performance was enhanced, and the final segmentation map F was obtained through $F_{body} + F_{edge}$.

3.4. Multi-Target Domain Adversarial Learning Strategy

During the multi-target domain adversarial learning strategy, in order to ensure relatively close feature distributions after spanning different domains and to better achieve transfer alignment between source and target domains, this paper added the adversarial loss terms $L_{S\to T_d}$ and $L_{S\to T_d}$ to the outputs of the daytime domain T_d and the nighttime domain T_n, respectively. Both discriminators had identical structures, weights, and training protocols, where the identification source domain image was 1, and the target domain image was 0. The binary cross-loss function [42] was utilized to make both $\tilde{F}_{T_d}$ and $\tilde{F}_{T_n}$ close to $\tilde{F}_S$. The antagonistic loss is defined as:

$$L_{adv} = L_{S\to T_d}(X_S, X_{T_d}) + L_{S\to T_n}(X_S, X_{T_n}), \tag{12}$$

In the experiments, we trained the generator and the discriminators alternately. The generator used in the source domain G_{pre} was pre-trained, and the target domain G_{seg} was transfered. The objective functions of $D_{S\to T_d}$ and $D_{S\to T_n}$ are defined as:

$$\begin{aligned} L_{S\to T_d}(X_S, X_{T_d}) = \min_{G_{seg}} \max_{D_{S\to T_d}} (&E_{X_S\sim p_{data}(X_S)}(\log D_{S\to T_d}(G_{pre}(X_S))) + \\ &E_{X_{T_d}\sim p_{data}(X_{T_d})}(1 - \log D_{S\to T_d}(G_{seg}(X_{T_d})))), \end{aligned} \tag{13}$$

$$\begin{aligned} L_{S\to T_n}(X_S, X_{T_n}) = \min_{G_{seg}} \max_{D_{S\to T_n}} (&E_{X_S\sim p_{data}(X_S)}(\log D_{S\to T_n}(G_{pre}(X_S))) + \\ &E_{X_{T_n}\sim p_{data}(X_{T_n})}(1 - \log D_{S\to T_n}(G_{seg}(X_{T_n})))), \end{aligned} \tag{14}$$

We used cross-entropy loss to train the semantic segmentation loss of the source domain. Moreover, we introduced the small pixel reweighting w_k to address the small target category imbalance, as shown in Equation (15).

$$L_{seg} = -\frac{1}{NC}\sum_{t=1}^{H\times W}\sum_{k=1}^{C} ||w_k GT^{(t,k)} \cdot log(\hat{Y}_S^{(t,k)})||_1, \tag{15}$$

where N is the total number of image pixels, k denotes category, $||\cdot||_1$ is the L_1 norm that sums up all the pixels, w_k is the pixel weight, $\hat{Y}_S^{(k)}$ is the prediction map $\hat{Y}_S$ from the kth channel of the source domain image obtained from the activation network N_Y, i.e., $\hat{Y} = N_Y(\tilde{F})$, and $GT^{(k)}$ is the ground truth of the kth category of the one-hot encoding. Specifically, for each category k, we first defined a weight $w'_k = -\log(p_k)$, where p_k denotes the percentage of all valid pixels that are labeled as category k in the source domain. Then, w_k was further normalized by $w_k = ((w'_k - \overline{w})/\theta_k) \cdot std + avg$, where $\overline{w}$ and θ_k are the mean and standard deviation of w'_k, respectively, and std and avg are preset constants to limit the value of w_k to positive. Finally, w_k was multiplied by the corresponding category channel in $\tilde{F}$ to generate the weighted probability map, and then the segmentation result was yielded via N_Y, as shown in Equation (16).

$$\hat{Y}^{(k)} = N_Y(w_k \cdot \tilde{F}^{(k)}), \tag{16}$$

where $\hat{Y}^{(k)} \in \mathbb{R}^{H\times W\times C}$.

Therefore, the total loss of the whole network is:

$$L_{total} = L_{en} + L_{seg} + L_{de} + L_{adv}. \tag{17}$$

In summary, we designed a segmentation network for decoupling body and edge. It predicted the body and edge features of the input image, applied L_1 and L_2 losses to constrain them, respectively, and was then synthesized into a segmentation feature map. After that, two discriminators were used to distinguish different domain outputs between source and daytime image and source and nighttime image in a multi-objective domain adversarial learning strategy. Additionally, probabilistic reweighting was used to optimize the segmentation prediction for small targets.

4. Experiments

4.1. Experimental Settings

To assess the performance of the proposed DLA-Net and its components, the Mean Intersection-to-Noise Ratio (mIoU), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) were employed as the evaluation metrics in the experiments and compared with the advanced methods. The mIoU is a widely used metric for evaluating the accuracy of pixel-level semantic segmentation models which calculates the ratio of intersection and union between predicted segmentation results and true labels. PSNR and the SSIM are commonly used metrics for evaluating image enhancement work. Both metrics compare the differences between the original and compressed/distorted images. PSNR measures image quality by comparing the Peak Signal-to-Noise Ratio, while the SSIM evaluates it in terms of structure, brightness, and contrast similarity. In addition, the following datasets were used for the training of all segmentation models and performance evaluation during the daytime–nighttime domain adaptive transfer process.

CityScapes [25]: The CityScapes dataset comprises 5000 street view images with a resolution of 2048 × 1024 divided into 2975 training images, 500 validation images, and 1525 testing images. Each image is annotated at the pixel level with 19 categories. We used the CityScapes training images as the training data in the training phase and the comparative experimental dataset for the decoupled body and edge segmentation modules;
Dark Zurich [5]: The Dark Zurich dataset comprises 2416 nighttime images, 2920 twilight images, and 3041 daytime images for training, all with a resolution of 1920 × 1080. The images in these three domains are roughly aligned using GPS localization of neighboring locations and panning/zooming operations in all directions. In this paper, 2416 nighttime images were utilized to train the network model (without utilizing twilight images). In addition to the above images used for training, Dark Zurich contains 201 annotated nighttime images, of which 50 were used for validation (Dark Zurich-val) and 151 for testing (Dark Zurich-test) and evaluation;
Nighttime Driving [10]: In the experiments, we exclusively utilized the Nighttime Driving test set, which comprises 50 nighttime images with a resolution of 1920 × 1080. All images are pixel-level annotated using 19 cityscape categories;
SICE dataset [43]: The Part 2 subset of the SICE dataset was utilized in this paper, which comprises of 229 multi-exposure sequences and the reference image corresponding to each sequence. In the experiments, only low-light images from the Part 2 subset were used.

In this study, we implemented the proposed adversarial learning domain adaptive semantic segmentation network using PyTorch on a single NVIDIA 3060 GPU, and all networks were trained using the same settings. Following [44], we trained the networks using an SGD optimizer and set the SGD optimizer momentum to 0.9 and a decay of 5×10^{-4}. The base learning rate of the network was 2.5×10^{-4}, and then the learning rate was reduced using a polynomial learning rate strategy with a decay power of 0.9. The batch size was 2. We used an Adam optimizer [45] to train the discriminators with the β set to (0.9, 0.99). The learning rate of discriminators followed the same decay strategy as the generator. The total enhancement loss L_{en} incorporates weights λ_1 and λ_2, which are selected from the intervals $[0.1 - 1.0]$ and $[20 - 25]$, respectively. These values were chosen based on previous similar work and different loss characteristics. After experimentation on the validation set, λ_1 and λ_2 were set to 0.4 and 20, respectively. In L_{de}, the weight of L_{edge} is determined by λ_3. This hyperparameter is set because L_{body} and L_{edge} are complementary, with fewer pixels in

the edge part relative to the body part. The default value of λ_3 is 1, but, in our experiments, we found that a value of 0.4 resulted in the best segmentation. To ensure the positivity of the values of w_k, we set $std = 0.05, avg = 1.0$ in the experiments. ResNet-101 [46] was used as the backbone. To facilitate smoother convergence during training, we used a total of 180,000 pre-training epochs on the CityScapes dataset with three different semantic segmentation models. Table 1 presents the performance of the three distinct semantic segmentation models on the validation sets of CityScapes and Dark Zurich.

Table 1. The performance of three distinct semantic segmentation models on the validation sets of CityScapes and Dark Zurich.

Method	Dark Zurich-Val	CityScapes-Val
RefineNet [47]	14.46	64.50
PSPNet [48]	11.44	64.97
DeepLab-v3+ [49]	11.58	63.77

4.2. Comparison with Other Methods

Comparison on Dark Zurich-test: In this paper, we first compare the proposed DLA-Net with several state-of-the-art methods on Dark Zurich-test [5], including CPSL [50], ProCA [51], and DiGA [52], as well as some other domain adaptation methods [6,22,53]. The performance results are summarized in Table 2.

Table 2. Results of the current state-of-the-art method and the DLA-Net proposed in this paper for each category in the Dark Zurich test set. CityScapes→ DZ-night denotes the adaptation from CityScapes to Dark Zurich-night. Bold font indicates the **best**, and underlining indicates the second-best.

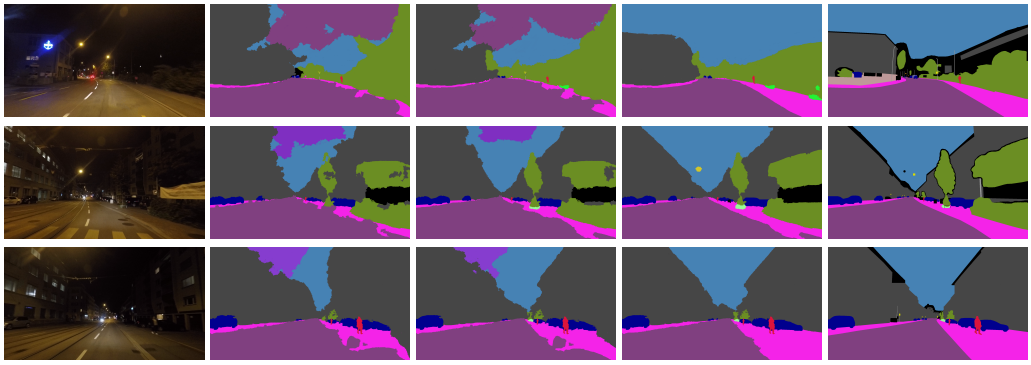
Method	Road	Sidewalk	Building	Wall	Fence	Traffic Light	Traffic Sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motocycle	Bicycle	Pole	mIou
RefineNet-CityScapes [47]	68.8	23.2	46.8	20.8	12.6	30.4	26.9	43.1	14.3	0.3	36.9	49.7	63.6	6.8	0.2	24.0	33.6	9.3	29.8	28.5
PSPNet-CityScapes [48]	79.0	21.8	53.0	13.8	11.2	20.2	21.9	43.5	10.4	20.2	37.4	33.8	64.1	6.4	0.0	52.3	30.4	7.4	22.5	28.8
DeepLab-v3+-CityScapes [49]	78.2	19.0	51.2	15.5	10.6	28.9	22.0	56.7	13.3	20.8	38.2	21.8	52.1	1.6	0.0	53.2	23.2	10.7	30.3	28.8
AdaptSegNet-CityScapes→ DZ-night [6]	86.1	44.2	55.1	22.2	4.8	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	22.1	30.4
ADVENT-CityScapes→DZ-night [53]	85.8	37.9	55.5	27.7	14.5	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	23.1	29.7
BDL-CityScapes→ DZ-night [22]	85.3	41.1	61.9	32.7	17.4	11.4	21.3	29.4	8.9	1.1	37.4	22.1	63.2	28.2	0.0	47.7	39.4	15.7	20.6	30.8
CPSL [50]	75.0	28.6	48.1	20.8	13.8	36.3	29.4	48.9	13.3	0.4	42.8	49.7	68.9	17.9	0.0	27.1	34.4	11.4	33.8	31.6
ProCA [51]	81.2	46.4	58.3	21.5	19.5	**40.0**	**41.1**	64.3	30.5	31.6	**53.0**	47.0	**75.0**	38.7	0.0	49.1	30.2	20.5	40.7	41.5
DiGA [52]	79.8	48.8	65.7	7.3	10.5	38.4	38.4	63.6	17.5	55.3	51.6	**53.0**	74.2	**62.0**	0.0	37.0	28.6	22.0	**40.6**	42.1
DLA-Net (RefineNet)	88.5	53.3	69.7	33.9	19.9	31.4	35.8	69.4	**32.1**	82.2	44.1	43.6	54.0	21.9	0.0	40.8	35.9	**24.0**	24.9	42.4
DLA-Net (PSPNet)	89.2	53.0	**74.0**	**40.2**	20.3	26.0	29.4	**71.2**	25.4	**83.2**	46.2	33.1	67.4	18.2	**0.3**	65.6	37.5	22.8	24.2	43.5
DLA-Net (DeepLab-v3+)	**89.5**	**59.2**	70.1	32.7	**22.0**	33.4	32.8	69.6	30.9	79.3	44.8	40.7	66.5	15.9	0.1	**72.1**	30.7	22.0	29.7	**44.3**

In Table 2, CPSL, ProCA, and DiGA are shown to have utilized the same baseline RefineNet, while the other methods employed DeepLab v3+. Additionally, all methods utilized ResNet-101 as the backbone [46], and the experimental dataset was Dark Zurich-test. DLA-Net with DeepLab-v3+, RefineNet, or PSPNet achieved superior or equivalent performance compared to existing methods on this dataset. It attained an overall improvement in mIoU of 2.2% compared to the highest score obtained by existing methods (DiGA). Furthermore, the DLA-Net proposed in this paper excelled in various categories, such as roads, sidewalks, and sky. For example, in the sky category, DLA-Net outperformed ProCA and DiGA by 51.6 mIoU and 27.9 mIoU, respectively, demonstrating its ability to accurately segment these categories despite a large daytime–nighttime domain gap. Figure 4 provides the visualization results of the comparison experiments with ProCA [51] and DiGA [52], highlighting the superior performance of DLA-Net in the categories of sky, road, and sidewalk.

Comparison on Nighttime Driving: We compared the proposed DLA-Net with some other baseline methods on Nighttime Driving test [10], and results are reported in Table 3.

Table 3. Comparison results of the proposed DLA-Net and some baseline methods on Nighttime Driving test. Bold font indicates the **best**, and underlining indicates the second-best.

Method	mIoU
RefineNet-CityScapes [47]	32.75
PSPNet-CityScapes [48]	25.44
DeepLab-v3+-CityScapes [49]	27.65
AdaptSegNet-CityScapes→ DZ-night [6]	34.5
ADVENT-CityScapes→DZ-night [53]	34.7
BDL-CityScapes→DZ-night [22]	34.7
CPSL [50]	38.2
ProCA [51]	46.7
DiGA [52]	**49.9**
DLA-Net (RefineNet)	43.82
DLA-Net (PSPNet)	44.59
DLA-Net (DeepLab-v3+)	47.08



(**a**) Input image. (**b**) ProCA. (**c**) DiGA. (**d**) DLA-Net. (**e**) Semantic GT.

Figure 4. Visualization results of DLA-Net and some other baseline methods on Dark Zurich-val.

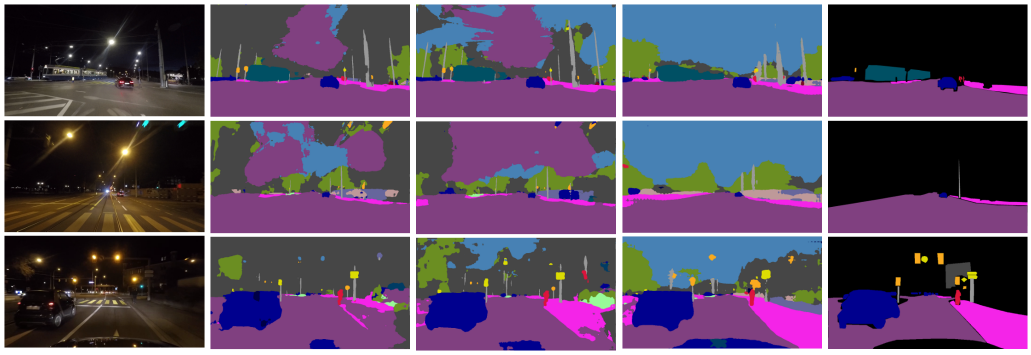
It is important to note that the Nighttime Driving test dataset is not as finely labeled as the Dark Zurich test dataset as some elements like buildings and vegetation are not labeled. Of the 50 images in the Nighttime Driving test dataset, only two are labeled with the sky category. Despite the limited number of labeled categories and the small dataset size, the DLA-Net with DeepLab-v3+ still achieved the second-best performance (DiGA was the top performer) on this dataset. Figure 5 reports the visualization results of the ProCA [51] and DiGA [52] comparison experiments. This underscores that the DLA-Net proposed in this paper can produce superior results even when working with a small number of samples and labeled categories for segmentation tasks.

Comparison of decoupling body and edge segmentation module with other advanced segmentation methods: This paper uses ResNet-101 as a backbone [46] on the CityScapes dataset [25] to compare the decoupling body and edge segmentation module with some state-of-the-art techniques. The experimental results are listed in Table 4.

As shown in Table 4, the decoupling body and edge segmentation method proposed in this paper achieved the highest mIoU among all methods, reaching 83.1 mIoU. This demonstrates the effectiveness of the body generation branch and the edge preservation branch in segmentation. The body branch with L_2 loss constraints obtained prominent body features, while the edge branch with L_1 loss constraints captured clear edge features. The combination of these features resulted in an overall segmentation map that has been experimentally proven to yield better segmentation performance. Ablation experiments between different components are discussed in the next section.

Table 4. Comparison of decoupled body and edge segmentation module with other advanced segmentation methods. Bold font indicates the **best**, and underlining indicates the second-best.

Method	Backbone	mIoU
DFN [54]	ResNet-101	79.3
PSANet [55]	ResNet-101	80.1
DenseASPP [56]	DenseNet-161	80.6
DANet [57]	ResNet-101	81.5
CCNet [58]	ResNet-101	81.4
BAFNet [59]	ResNet-101	81.4
ACFNet [60]	ResNet-101	81.9
GFFnet [61]	ResNet-101	82.3
X. Li et al. [62]	ResNet-101	82.8
Ours	ResNet-101	**83.1**



(**a**) Input image. (**b**) ProCA. (**c**) DiGA. (**d**) DLA-Net. (**e**) Semantic GT.

Figure 5. Visualization results of ablation experiments with different loss functions.

Comparison of low-light image enhancement sub-network with other methods: In this paper, reference image quality assessment metrics PSNR and SSIM were used to quantitatively compare the performance of different methods on the SICE Part 2 test set [43]. Higher values of SSIM and PSNR indicate that the enhanced image is closer to the ground truth in terms of structural properties and pixel-level image content, respectively. The experimental results are presented in Table 5.

Table 5. Comparison of low-light image enhancement sub-networks with other methods. Bold font indicates the **best**, and underlining indicates the second-best.

Method	PSNR↑	SSIM↑
MBLLEN [34]	14.78	0.534
RetiexNet [63]	15.56	0.525
RUAS [64]	16.40	0.500
ZeroDCE [37]	14.86	0.559
SCI [65]	14.78	0.522
EnlightenGAN [42]	17.48	**0.651**
Ours	**18.10**	0.638

Table 5 reveals that, despite not using any paired or unpaired training data, the LIE-SubNet proposed in this paper still achieved the best PSNR and second-best SSIM results (EnlightenGAN was the top performer). Combining the mapping curve and the depth network resulted in a 1.1% improvement compared to the second-best performance, and the mapping curve with multiple iterations made the overall pixels of the low-light images more uniform (see Figure 5).

In this subsection, we present comparative experiments of the overall method using the Dark Zurich dataset and the Nighttime Driving dataset. The experimental results validate the excellent performance of the proposed method. However, it should be noted that the Dark Zurich dataset mostly contains unobstructed image foregrounds, and there exists a one-to-one correspondence between the daytime and nighttime images. As a result, DLA-Net is able to perform optimally and obtain excellent results. When there is occlusion in the foreground of an image, LIE-SubNet and decoupled subject and segmentation networks may not be effective in enhancing the image and performing decoupled segmentation.

4.3. Ablation Study

To demonstrate the effectiveness of the different components of the proposed DLA-Net in this paper, several ablation experiments were conducted on several model variants. The results of the ablation experiments on different components are detailed below.

Ablation study on decoupling body and edge modules: The effectiveness of the two branches in the decoupling body and edge segmentation network is illustrated in Table 6, where ρ and δ denote the body generation branch and the edge preservation branch, respectively. The direct addition of the body generation and edge preservation branches in DeepLab-V3+ [49] improved the segmentation effect by 1.7%, implying that both branches are effective. After adding L_{body} and L_{edge}, respectively, there were further improvements of 0.5% and 0.4% in performance, demonstrating that using L_2 and L_1 loss constraints can facilitate the model's learning of different features. Finally, when all losses were combined, the performance was further improved by 1.7%. This paper also investigated the necessity of the F_{detail} module, and its removal resulted in a decrease in segmentation performance of about 0.7%.

Table 6. Comparison of decoupled body and edge segmentation module with other advanced segmentation methods. ✓ indicates that L_{body} or L_{edge} was used.

Method	L_{body}	L_{edge}	mIoU	Δ(%)
DeepLab-v3+ [49]			74.6	-
+ρ & δ	-	-	75.9	1.7↑
	✓	-	76.3	0.5↑
	-	✓	76.6	0.4↑
	✓	✓	77.6	1.7↑
w/o F_{detail}	✓	✓	77.1	0.7↓

The results of the ablation studies for each component of the decoupling body and edge segmentation module are shown in Table 7. After removing the average pooling layer and the encoder–decoder in the body generation branch, the model performance decreases by 1.5% and 1.0% accordingly. After removing the edge preservation branch, the model performance decreased by 0.4%. Therefore, removing the three modules individually leads to varying magnitudes of degradation in segmentation network performance. This indicates that average pooling and codecs help predict the body feature in the body generation branch and that average pooling improves the performance to a greater extent, while the entire edge preservation branch can also elevate the performance of the segmentation network.

Table 7. Ablation study on effect of each component.

Method	mIoU	Δ(%)
DeepLab-v3+ [49] +ρ & δ	77.6	-
w/o ρ average pooling	76.4	1.5 ↓
w/o ρ encoder–decoder	76.8	1.0 ↓
w/o δ	77.2	0.4 ↓

Ablation study for each loss in the low-light image enhancement sub-network: The visualization results of the ablation experiments with different loss functions in the LIE-SubNet are presented in Figure 6. After removing the exposure control loss L_{ec}, the image exhibited overexposure in areas with strong lighting, underscoring the effectiveness of exposure constraints in the network. The removal of the color consistency loss L_{cc} resulted in the overall image's severe color deviation. With the removal of the light smoothing loss L_{is}, artifacts appeared between adjacent regions in the image. These experiments highlight the critical contributions of each loss function used in this paper in the LIE-SubNet.

(**a**) Input image. (**b**) LIE-SubNet. (**c**) w/oL_{ec}. (**d**) w/oL_{cc}. (**e**) w/oL_{is}.

Figure 6. Visualization results of ablation experiments with different loss functions.

Ablation study on different components of DLA-Net: As shown in Table 8, AdaptSegNet [6] was used as the baseline and DLA-Net as the full model. It was observed that, although X_{T_d} was unlabeled, using roughly aligned X_{T_d} to predict X_{T_n} was quite important and also played a key role in DLA-Net. It reduced the segmentation results by 36.8% without using X_{T_d}, indicating that the training in the daytime domain is quite critical in the network. The LIE-SubNet and the corresponding loss function L_{en} also contribute to the whole network. Meanwhile, the utilization of the decoupling body and edge loss L_{de} in this paper yielded superior results when compared to applying the cross-entropy loss directly to computing the segmentation loss. The performance disparity between the two methods is notable, and not using the L_{de} loss outperformed using the cross-entropy loss by 38%. In addition, the adoption of probabilistic reweighting in experiments enhanced the segmentation performance, affirming its effectiveness as an auxiliary tool.

Table 8. Ablation study of several DLA-Net (DeepLab-v3+) modules proposed in this paper on Dark Zurich-val.

Method	mIoU
ProCA [51]	25.47
DiGA [52]	25.21
AdaptSegNet-CityScapes→ DZ-night [6]	19.13
w/o X_{T_d}	22.58
w/o LIE-SubNet & L_{en}	32.45
w/o L_{en}	33.86
w/o L_{de}	20.19
w cross-entropy loss in L_{de}	32.96
w/o probability reweighting	31.68
w/o pre-trained segmentation model	29.78
DLA-Net	35.74

This chapter presents comparison and ablation experiments on DLA-Net and its components across multiple datasets. The results demonstrate that DLA-Net efficiently segments images in nighttime scenes without using labeled images or synthetic datasets. LIE-SubNet effectively brightens low-light images, and the decoupling body and edge segmentation effectively predict feature maps with a uniform body and clear edge. However, DLA-Net struggles with domain adaptation when faced with large domain gaps caused by differences in styles and inherent variations between datasets, such as in urban street scenes. Future work will focus on conducting in-depth research to better understand these differences and adapt to a wider range of scenes and datasets.

5. Conclusions

In this paper, we proposed the DLA-Net, an adversarial learning domain adaptive semantic segmentation network capable of performing semantic segmentation of nighttime low-light images. DLA-Net leverages a combination of mapping curve iteration and a deep network to enhance low-light images, ensuring their distributions align with those from different domains. The segmentation network employs the decoupling body and edge modules that can efficiently obtain body and edge features, respectively. After the segmentation network, two discriminators are used to differentiate outputs from different domains. Therefore, a multi-target domain adversarial learning strategy is constituted between the generator and the discriminators to realize the adversarial learning domain adaption for multi-target domains. The experimental results underscore the efficacy of each designed component, showcasing outstanding performance on datasets such as Dark Zurich and Nighttime Driving. State-of-the-art performance is also obtained on unlabeled or thin labeled datasets, and segmentation performance is better on recognizable classes with large domain gaps. However, DLA-Net does not perform well in adapting to the different styles and inherent differences between datasets. Future work will investigate how to adapt to more scenarios and datasets.

Author Contributions: Conceptualization, M.W. and Z.Z.; methodology, M.W. and Z.Z.; software, M.W. and Z.Z.; validation, M.W., Z.Z. and H.L.; formal analysis, M.W. and Z.Z.; investigation, M.W. and Z.Z.; resources, M.W. and Z.Z.; data curation, M.W., Z.Z. and H.L.; Writing—original draft, M.W. and Z.Z.; Writing—review and editing, M.W., Z.Z. and H.L.; visualization, Z.Z.; supervision, M.W. and H.L.; project administration, M.W.; funding acquisition, M.W. and H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (62062048) and Yunnan Provincial Science and Technology Plan Project (202201AT070113). This work is also supported by the Faculty of Information Engineering and Automation, Kunming University of Science and Technology.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in CityScapes at https://www.cityscapes-dataset.com/, Dark Zurich at https://trace.ethz.ch/publications/2019/GCMA_UIoU/, Nighttime Driving at http://people.ee.ethz.ch/~daid/NightDriving/ and SICE dataset at https://github.com/csjcai/SICE?tab=readme-ov-file.

Conflicts of Interest: The authors declare no conflicts of interest.

References

2. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
2. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
3. Chen, X.; Williams, B.M.; Vallabhaneni, S.R.; Czanner, G.; Williams, R.; Zheng, Y. Learning active contour models for medical image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11632–11640.
4. Zhang, X.; Chen, Y.; Zhu, B.; Wang, J.; Tang, M. Part-aware context network for human parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8971–8980.
5. Sakaridis, C.; Dai, D.; Gool, L.V. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7374–7383.
6. Tsai, Y.H.; Hung, W.C.; Schulter, S.; Sohn, K.; Yang, M.H.; Chandraker, M. Learning to adapt structured output space for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7472–7481.
7. Bang, G.; Lee, J.; Endo, Y.; Nishimori, T.; Nakao, K.; Kamijo, S. Semantic and Geometric-Aware Day-to-Night Image Translation Network. *Sensors* **2024**, *24*, 1339. [CrossRef] [PubMed]

8. Manettas, C.; Nikolakis, N.; Alexopoulos, K. Synthetic datasets for Deep Learning in computer-vision assisted tasks in manufacturing. *Procedia CIRP* **2021**, *103*, 237–242. [CrossRef]
9. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423.
10. Dai, D.; Van Gool, L. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 3819–3824.
11. Sakaridis, C.; Dai, D.; Van Gool, L. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 3139–3153. [CrossRef] [PubMed]
12. Kurmi, V.K.; Bajaj, V.; Subramanian, V.K.; Namboodiri, V.P. Curriculum based dropout discriminator for domain adaptation. *arXiv* **2019**, arXiv:1907.10628.
13. Sun, L.; Wang, K.; Yang, K.; Xiang, K. See clearer at night: towards robust nighttime semantic segmentation through day-night image conversion. In *Artificial Intelligence and Machine Learning in Defense Applications*; SPIE: Strasbourg, France, 2019; Volume 11169, pp. 77–89.
14. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
15. Hung, W.C.; Tsai, Y.H.; Liou, Y.T.; Lin, Y.Y.; Yang, M.H. Adversarial learning for semi-supervised semantic segmentation. *arXiv* **2018**, arXiv:1802.07934.
16. Nag, S.; Adak, S.; Das, S. What's there in the dark. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 2996–3000.
17. Fan, R.; Xie, J.; Yang, J.; Hong, Z.; Xu, Y.; Hou, H. Multiscale Change Detection Domain Adaptation Model Based on Illumination–Reflection Decoupling. *Remote Sens.* **2024**, *16*, 799. [CrossRef]
18. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
19. Wu, X.; Wu, Z.; Guo, H.; Ju, L.; Wang, S. 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, Nashville, TN, USA, 20–25 June 2021; pp. 15769–15778.
20. Hoffman, J.; Wang, D.; Yu, F.; Darrell, T. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv* **2016**, arXiv:1612.02649.
21. Pathak, D.; Krahenbuhl, P.; Darrell, T. Constrained convolutional neural networks for weakly supervised segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1796–1804.
22. Li, Y.; Yuan, L.; Vasconcelos, N. Bidirectional learning for domain adaptation of semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6936–6945.
23. Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.Y.; Isola, P.; Saenko, K.; Efros, A.; Darrell, T. Cycada: Cycle-consistent adversarial domain adaptation. In Proceedings of the International Conference on Machine Learning, Chengdu, China, 15–18 July 2018; pp. 1989–1998.
24. Wu, Z.; Han, X.; Lin, Y.L.; Uzunbas, M.G.; Goldstein, T.; Lim, S.N.; Davis, L.S. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 518–534.
25. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
26. Liu, M.Y.; Breuel, T.; Kautz, J. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*; MIT Press: Long Beach, CA, USA, 2017; Volume 30.
27. Xie, Q.; Luong, M.T.; Hovy, E.; Le, Q.V. Self-training with noisy student improves imagenet classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10687–10698.
28. Lian, Q.; Lv, F.; Duan, L.; Gong, B. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6758–6767.
29. Zhang, Y.; David, P.; Gong, B. Curriculum domain adaptation for semantic segmentation of urban scenes. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2020–2030.
30. Anoosheh, A.; Sattler, T.; Timofte, R.; Pollefeys, M.; Van Gool, L. Night-to-day image translation for retrieval-based localization. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May2019; pp. 5958–5964.
31. Vertens, J.; Zürn, J.; Burgard, W. Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October–24 January 2021; pp. 8461–8468.
32. Di, S.; Feng, Q.; Li, C.G.; Zhang, M.; Zhang, H.; Elezovikj, S.; Tan, C.C.; Ling, H. Rainy night scene understanding with near scene semantic adaptation. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 1594–1602. [CrossRef]

33. Valada, A.; Vertens, J.; Dhall, A.; Burgard, W. Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 4644–4651.
34. Lv, F.; Lu, F.; Wu, J.; Lim, C. MBLLEN: Low-Light Image/Video Enhancement Using CNNs. In Proceedings of the BMVC, Newcastle, UK, 3–6 September 2018; Volume 220, p. 4.
35. Bychkovsky, V.; Paris, S.; Chan, E.; Durand, F. Learning photographic global tonal adjustment with a database of input/output image pairs. In Proceedings of the CVPR 2011, Providence, RI, USA, 20–25 June 2011; pp. 97–104.
36. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2599–2613. [CrossRef] [PubMed]
37. Li, C.; Guo, C.; Loy, C.C. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4225–4238. [CrossRef] [PubMed]
38. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. In *Advances in Neural Information Processing Systems*; Google DeepMind: London, UK, 2015; Volume 28.
39. Zhu, X.; Xiong, Y.; Dai, J.; Yuan, L.; Wei, Y. Deep feature flow for video recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2349–2358.
40. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2; pp. 1735–1742.
41. Li, X.; Yang, Y.; Zhao, Q.; Shen, T.; Lin, Z.; Liu, H. Spatial pyramid based graph reasoning for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8950–8959.
42. Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; Wang, Z. Enlightengan: Deep light enhancement without paired supervision. *IEEE Trans. Image Process.* **2021**, *30*, 2340–2349. [CrossRef] [PubMed]
43. Chen, Y.S.; Wang, Y.C.; Kao, M.H.; Chuang, Y.Y. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6306–6314.
44. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
45. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
47. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
48. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
49. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
50. Li, R.; Li, S.; He, C.; Zhang, Y.; Jia, X.; Zhang, L. Class-balanced pixel-level self-labeling for domain adaptive semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11593–11603.
51. Zhang, P.; Zhang, B.; Zhang, T.; Chen, D.; Wang, Y.; Wen, F. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12414–12424.
52. Shen, F.; Gurram, A.; Liu, Z.; Wang, H.; Knoll, A. DiGA: Distil to Generalize and then Adapt for Domain Adaptive Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 15866–15877.
53. Vu, T.H.; Jain, H.; Bucher, M.; Cord, M.; Pérez, P. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2517–2526.
54. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Learning a discriminative feature network for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1857–1866.
55. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.C.; Lin, D.; Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 267–283.
56. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3684–3692.

57. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
58. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612.
59. Ding, H.; Jiang, X.; Liu, A.Q.; Thalmann, N.M.; Wang, G. Boundary-aware feature propagation for scene segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6819–6829.
60. Zhang, F.; Chen, Y.; Li, Z.; Hong, Z.; Liu, J.; Ma, F.; Han, J.; Ding, E. Acfnet: Attentional class feature network for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6798–6807.
61. Li, X.; Zhao, H.; Han, L.; Tong, Y.; Yang, K. Gff: Gated fully fusion for semantic segmentation. *arXiv* **2019**, arXiv:1904.01803.
62. Li, X.; Li, X.; Zhang, L.; Cheng, G.; Shi, J.; Lin, Z.; Tan, S.; Tong, Y. Improving semantic segmentation via decoupled body and edge supervision. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XVII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 435–452.
63. Wei, C.; Wang, W.; Yang, W.; Liu, J. Deep retinex decomposition for low-light enhancement. *arXiv* **2018**, arXiv:1808.04560.
64. Liu, R.; Ma, L.; Zhang, J.; Fan, X.; Luo, Z. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10561–10570.
65. Ma, L.; Ma, T.; Liu, R.; Fan, X.; Luo, Z. Toward fast, flexible, and robust low-light image enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5637–5646.