Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.052285

ARTICLE

ED-Ged: Nighttime Image Semantic Segmentation Based on Enhanced Detail and Bidirectional Guidance

Xiaoli Yuan, Jianxun Zhang*, Xuejie Wang and Zhuhong Chu

College of Computer Science and Engineering, Chongqing University of Technology, Chongqing, 400054, China

*Corresponding Author: Jianxun Zhang. Email: zjx@cqut.edu.cn

Received: 28 March 2024 Accepted: 27 June 2024 Published: 15 August 2024

ABSTRACT

Semantic segmentation of driving scene images is crucial for autonomous driving. While deep learning technology has significantly improved daytime image semantic segmentation, nighttime images pose challenges due to factors like poor lighting and overexposure, making it difficult to recognize small objects. To address this, we propose an Image Adaptive Enhancement (IAEN) module comprising a parameter predictor (Edip), multiple image processing filters (Mdif), and a Detail Processing Module (DPM). Edip combines image processing filters to predict parameters like exposure and hue, optimizing image quality. We adopt a novel image encoder to enhance parameter prediction accuracy by enabling Edip to handle features at different scales. DPM strengthens overlooked image details, extending the IAEN module's functionality. After the segmentation network, we integrate a Depth Guided Filter (DGF) to refine segmentation outputs. The entire network is trained end-to-end, with segmentation results guiding parameter prediction optimization, promoting self-learning and network improvement. This lightweight and efficient network architecture is particularly suitable for addressing challenges in nighttime image segmentation. Extensive experiments validate significant performance improvements of our approach on the ACDC-night and Nightcity datasets.

KEYWORDS

Night driving; semantic segmentation; nighttime image processing; adverse illumination; differentiable filters

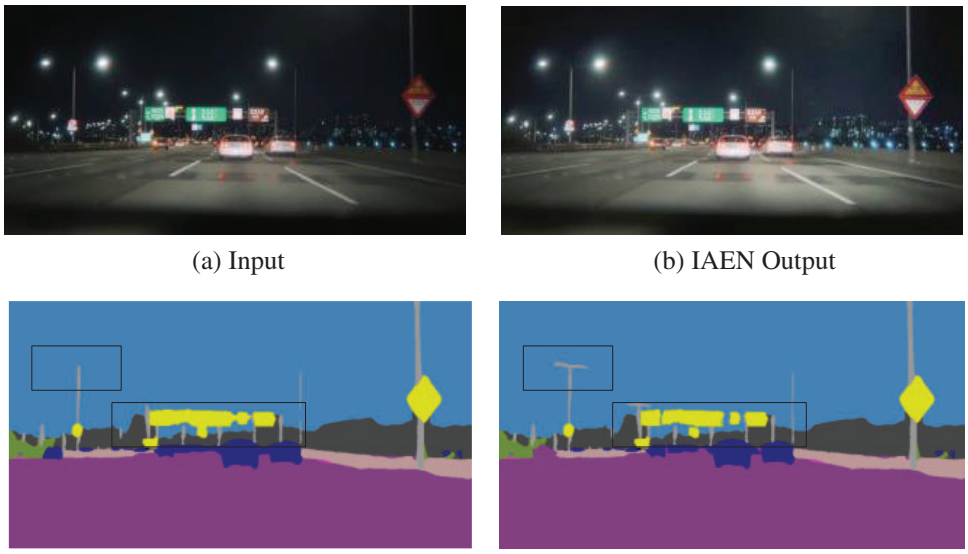
1 Introduction

Semantic segmentation is a crucial computer vision task with the primary goal of assigning corresponding labels to each pixel in an input image. This task plays a pivotal role in various practical applications, including but not limited to autonomous driving [1], medical image processing [2] and virtual reality [3]. As deep learning gains momentum, semantic segmentation techniques powered by deep neural networks have yielded outstanding performance on conventional datasets. Nevertheless, these methods commonly face challenges related to limited generalization when dealing with nighttime images. Therefore, the main purpose of this study is to provide a more effective solution to the semantic segmentation problem in nighttime environments, thus promoting the development of computer vision technology in nighttime applications.

Nighttime image segmentation faces three major challenges. Firstly, acquiring large-scale nighttime image datasets becomes difficult due to visual perception differences. To address this issue, recently, numerous new datasets have surfaced, such as Nightcity [4] and ACDC [5]. Secondly, there is insufficient natural light in nighttime environments, while streetlights, car lights, and other light sources may introduce strong reflections and shadows. This makes it difficult for segmentation algorithms to recognize and segment. Thirdly, nighttime driving scenes exhibit a complex diversity, with significant variations in the shapes, colors, and sizes of road signs, traffic signals, vehicles, and pedestrians. These factors pose challenges that existing daytime segmentation methods find difficult to overcome. To address these difficulties, certain scholars suggest using domain adaptation strategies to shift semantic segmentation models into nighttime applications, without needing specific nighttime data annotations. For instance, DANNet [6] introduced a domain adaptation network based on adversarial learning, addressing nighttime challenges by incorporating an image-relighting subnetwork. Certain studies [7] utilized image transfer models to stylize nighttime or daytime images to construct datasets. However, style transfer networks struggle to fully exploit semantic information and increase inference time. Literature [8] devised an end-to-end framework focusing on nighttime semantic segmentation, proposing a cross-domain correlation distillation algorithm.

This paper introduces a novel approach for preprocessing nighttime images using a differentiable image processing module and a learnable guided filter to enhance segmentation results. Improving image quality is emphasized as crucial for better segmentation performance. The integration of a learnable guided filter allows continuous refinement of segmentation boundaries, closely matching ground truth values. Building on this, a gated differentiable image processing framework is developed, enabling convolutional neural networks to execute multiple image processing operations concurrently and learn the optimal weight combination for their outputs. This framework avoids issues associated with sequential processing and additional effects from pipeline processing, such as inaccuracies in predicting sharpening parameters after exposure adjustment. A multi-scale visual encoder further enhances segmentation performance by leveraging large-scale features for global information and small-scale features for local details. Comprehensive experiments on the Nightcity and ACDC datasets validate the remarkable performance enhancement of ED-Ged in nighttime semantic segmentation. Fig. 1 shows the visual output of each module, highlighting significant improvements in lighting, detail, and segmentation boundaries. Specifically, ED-Ged shows improvements of 3.23%, 4.21%, and 3.95% over baseline models, and 2.2% over state-of-the-art methods [9] on the Nightcity dataset. The main contributions of this paper can be summarized as follows:

1) We propose an innovative image enhancement module based on image-adaptive filtering, combining deep neural networks and gating mechanisms to enhance image quality comprehensively. Easily integrable into existing segmentation models, this module significantly improves nighttime image segmentation performance.
2) We introduce an innovative multi-scale adaptive feature extractor that predicts content-related weights for image-adaptive enhancement, progressively improving the input image.
3) We designed a detail processing module for small object detection with two branches: the contextual branch and the edge branch. The contextual branch captures global contextual cues to enhance image components, while the edge branch uses Sobel operators to detect edges and highlight texture features.
4) We designed learnable deep-guided filters to guide segmentation results and enable continuous learning, improving parameter prediction accuracy in the adaptive enhancement module and achieving clearer segmentation.



(a) Input (b) IAEN Output

(c) Segmentation Network Output (d) DGF Output

Figure 1: Visualization results of the output images from different modules

2 Related Work

2.1 Semantic Segmentation

Semantic segmentation finds wide applications in various domains such as autonomous driving [1], medical image processing [10] virtual reality [11] and indoor navigation [12]. With the advancement of deep learning, methods based on it have garnered attention and gradually replaced traditional algorithms [13,14]. Scholars continually refine and optimize these methods, proposing innovative models. FCN [15], U-Net [16] and variants of DeepLab [17,18] are noteworthy as they address different challenges in semantic segmentation. For instance, the foundational framework for employing deep learning techniques to tackle image semantic segmentation problems is established in reference [15]. With U-Net, based on the FCN principle, is widely applied in medical image segmentation. In 2017, PSPNet [19] and RefineNet [20] respectively tackled issues such as feature map resolution and multi-scale feature extraction. In 2023, Zou et al. [21] proposed cross domain road segmentation, greatly solved the road segmentation problem.

2.2 Image Adaptation

Image adaptation finds extensive application across various low-to-high level vision tasks [22,23]. For instance, classical approaches to image enhancement dynamically compute image transformation parameters based on feature-driven analysis. Reference [24] introduced the R2RNet framework, which jointly trains a nocturnal image enhancer alongside a depth estimator, offering an effective solution to regional underexposure issues encountered in nighttime scenes. Fu et al. [25] put forth a brightness optimization algorithm capable of fine-tuning amplification factors in response to variances found within each photograph's light dispersion attributes. Recently, Liu et al. [26] proposed using local feature frequency to address low-light enhancement, employing an entropy map to guide the convolution of different receptive fields. Reference [27] introduced a lightweight convolutional neural network (CNN), called IA-YOLO, to adaptively predict filter parameters, thereby achieving

superior detection performance. Subsequently, Liu et al. [9] extended IA-YOLO to nighttime semantic segmentation tasks, achieving promising results, segmentation task and obtained good results.

2.3 Nighttime Semantic Segmentation

Nighttime semantic segmentation [28] is crucial for safe autonomous driving. However, most existing models are trained and tested on daytime scenes with sufficient illumination, leading to a significant performance drop in nighttime scenarios. To address this issue, Zhang et al. [29] proposed a guided curriculum adaptation framework, effectively utilizing the correspondence between images from different time points. Some subsequent works, Vertens et al. [30] introduced thermal infrared images as supplementary inputs to RGB images to enhance nighttime semantic segmentation performance. Wang et al. [31] proposed SFNet-N, this work presents a cutting-edge nighttime segmentation system, proficient in discerning entities obscured by poorly lit circumstances, thus resolving problems stemming from indistinct edges due to minimal semantic differences typical of low-light image content. Literature [32] proposed a blurry information compensation strategy that combines generative models and fusion of intermediate layer outputs, aiming to enhance spatial semantics. Additionally, an irregular attention mechanism is integrated to precisely capture the contours of moving targets. Yang et al. [33] proposed a bidirectional fusion network, Bi-Mix, which mutually promotes image transformation and segmentation models through bidirectional training.

3 Proposed Method

3.1 Framework Overview

When targeting nighttime outdoor scenes, images captured by the camera may encounter issues such as underexposure and motion blur, resulting in overall low contrast, unclear details, and reduced visibility in the images. Given these limiting factors, current semantic segmentation models often fail to achieve accurate segmentation results when dealing with nighttime environments. To improve the accuracy of nighttime image segmentation and alleviate the difficulties associated with nighttime segmentation, our primary task is to address exposure discrepancies and ensure uniform lighting. Therefore, we introduce an Image Adaptive Enhancement Network (IAEN) before the segmentation network. As shown in Fig. 2, it consists of a visual encoder with five convolutional layers, a parameter predictor Edip, a Detail Processing Module (DPM), and multiple image processing filters (Mdif). Following the segmentation network, we combine the segmentation results with guidance images generated by a Depth Guided Filter (DGF) to further enhance the segmentation results using the guidance images. The entire network is trained end-to-end, where segmentation results continually guide the improvement process of parameter prediction, facilitating autonomous learning and performance enhancement of the network.

3.2 Network Architecture

Below, we will provide a detailed introduction to all modules of the network, including the visual encoder, parameter predictor Edip, Detail Processing Module (DPM), multiple image processing filters (Mdif), and Depth Guided Filter (DGF).

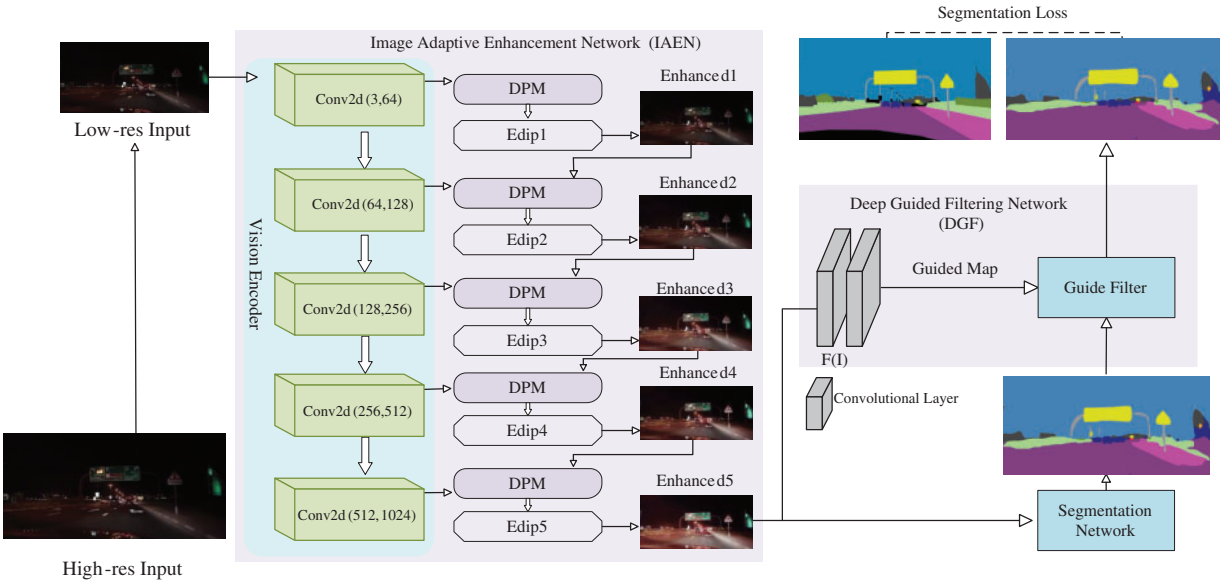


Figure 2: The overall framework of the ED-Ged model. The method first extracts image features from down sampled input images and passes them through an Image Adaptive Enhancement Network. Subsequently, the enhanced images are fed into a segmentation network for semantic segmentation, followed by post-processing with a Deep Guided Filter. The entire method is trained end-to-end, and subsequent ablation experiments will demonstrate the benefits of progressive image enhancement

3.2.1 Adaptive Image Enhancement Network

1) Gate-Based Parameter Predictor Edip

In image signal processing (ISP), adaptable filters enhance image quality but require manual parameter adjustment by experienced engineers, which is time consuming and costly. Standard parameters are hard to establish, as each image may need unique settings, particularly for nighttime image processing.

To better enhance the input image, we propose Edip, a multi-stage image enhancement technique based on a gating mechanism, as illustrated in Fig. 3. Edip consists of several gated image processing modules, called Ebip (where ip represents different image processing operations). performing individual image enhancement tasks. Outputs are fused using predicted weighting coefficients. Each Ebip integrates a linear component with differentiable image handling kernels, gating thresholds, and normalization operations. The linear layer computes parameters for the filters and a scalar value for gating inputs. A shared visual encoder extracts a common feature input for each Ebip module. Ultimately, the outputs of all Ebip modules are aggregated to form the input to Edip, thereby generating the enhanced image, represented mathematically:

$$z = N\left(\sum\nolimits_i \left(f_i\left(x\right)\right) * w_i\right) \tag{1}$$

where x is the input image captured in a nighttime environment, Z is the enhanced and clear image. $f_i(x)$ represents the $w_i \in [0, 1]$ operation weighted by individual scalar gates, and N is the min-max normalization operation. This normalization is employed to ensure that the pixel intensity ranges are consistent across all image processing operations.

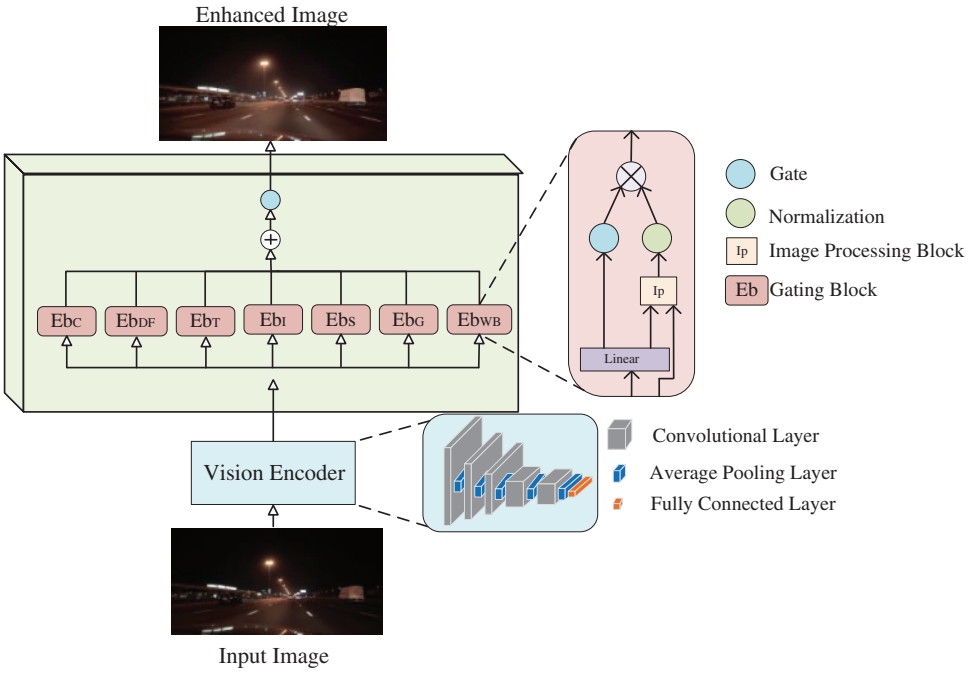


Figure 3: The structure of Edip, which includes five convolutional visual encoders, with each pink block representing a different image processing operation

It is worth noting that we introduce a gating mechanism for the parallel operation of image processing modules, controlled by corresponding gating thresholds with assigned weights. This enables multiple sets of image processing parameters to be extracted concurrently, improving parameter prediction accuracy compared to traditional serial processing methods. Additionally, our parameter predictor adopts a multi-level joint embedding strategy, where the output of each Edip module is progressively fed back to the next one, enhancing the image. This approach guides each Edip module using features from different levels of the visual encoder, thereby capturing a broader range of contextual information.

2) Visual Encoder

The proposed Edip block employs a five-layered convolutional visual encoder, starting with 64 channels and increasing to 1024 channels across five levels. After each convolution, average pooling with a kernel size of three and stride two is applied. A global pooling layer produces a 1×1024 output, followed by a fully connected layer projecting this output into a 256-dimensional space. The Edip block utilizes these 256-dimensional representations from different points in the visual encoder to compute image processing parameters and gate values, facilitating targeted image improvements.

3) Multiple Image Processing Filters

The proposed Mdif consists of 7 tunable differentiable filters, including Identity, Defog, White balance, Gamma, Contrast, Tone, and Sharpen. These filters can be utilized for image enhancement and denoising purposes. Specifically, the Identity filter retains the original information of the image, the Defog filter removes fog from the image, the White balance filter adjusts the color temperature of the image, the Gamma filter is used to adjust the brightness of the image, the Contrast filter

increases the image contrast, the Tone filter adjusts the image hue, and the Sharpen filter enhances the image sharpness. The hyperparameters of these filters are predicted by the parameter predictor mentioned earlier, and they can be adjusted according to different scenarios to achieve optimal image enhancement effects. The Defog filter is specifically designed for foggy weather conditions. White balance, Gamma, Contrast, and Tone can all be collectively referred to as per-pixel filters. Per-pixel filters refer to filters that map input pixel values $p_i = (r_i, g_i, b_i)$ to output pixel values $p_o = (r_o, g_o, b_o)$, where (r, g, b) denote the contributions of the red, green, and blue color streams, respectively. As shown in Table 1.

Table 1: Image Processing (IP) operations, further details on the equations can be found in reference [9]

Filter	Parameters	Mapping function
White balance	W_r, W_g, W_b	$I_{wb} = \left(W_r r_i, W_g g_i, W_b b_i\right)$
Gamma	β	$Ig = I^{\beta}$
Contrast	α	$I_{contrast} = \alpha \cdot En\,(I) + (1-\alpha) \cdot I$
Tone	I_i	$I_{tone} = \left(L_{tr}\,(r_i)\,, L_{tg}\,(g_i)\,, L_{tb}\,(b_i)\right)$
Identity	–	$I_{identity} = I$

Sharpen Filter: Image sharpening can accentuate image details. Nighttime images often suffer from blurriness and lack of detail due to lighting conditions. By applying a Sharpening filter, we can enhance the edges and details in the image, especially small objects like traffic signs, thereby improving the image clarity and visual appeal. The process of sharpening can be described as:

$$F(x, \gamma) = I(x) + \gamma\,(I(x) - Gau(I(x))) \tag{2}$$

where $I(x)$ represents the input image. $Gau(I(x))$ represents a Gaussian filter, γ is a positive scaling factor. The parameters x and γ possess differentiability. It is important to note that the degree of sharpening can be adjusted through optimization to achieve better segmentation performance.

Defog Filter: Analysis of datasets reveals that nighttime images and foggy images share similarities due to lighting constraints, causing blurriness and loss of detail. To address nighttime image issues, we use defogging, as our parameter predictor adjusts based on image characteristics. This defogging filter does not affect normal images. According to reference [34], the formation of foggy images can be represented as:

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{3}$$

where $I(x)$ symbolizes the foggy image, and $J(x)$ stands for the pristine version thereof. A embodies the all-pervading atmospheric illumination, whilst $t(x)$ delineates the media transmission matrix, formulated as follows:

$$t(x) = e^{-\beta} d(x) \tag{4}$$

where β represents the atmospheric scattering coefficient, and $d(x)$ is the scene depth. To restore a clean image $J(x)$, we shift the focus to obtaining the atmospheric light A and the transmission map $t(x)$. Based on Eq. (3), an approximate solution for $t(x)$ can be obtained [27]:

$$t(x) = 1 - \min_{C}\left(\min_{y \in \Omega(x)} \frac{I^c(y)}{A^C}\right) \tag{5}$$

To control the degree of defogging, we introduce a parameter w, which is also a hyperparameter that the parameter predictor needs to predict, specifically:

$$t(x,w) = 1 - w\min_{C}\left(\min_{y\in\Omega(x)}\frac{I^c(y)}{A^C}\right) \tag{6}$$

Since the operations are differentiable, we can optimize w through backpropagation, making the defogging filter more conducive to foggy image segmentation.

3.2.2 Detail Processing Module

To address the complex variability of nighttime images, relying solely on traditional image enhancement techniques often fails to achieve the desired results. Therefore, we propose a Detail Processing Module (DPM) aimed at enhancing the performance of the Adaptive Image Enhancement Network for better outcomes. As shown in Fig. 4, this module comprises a contextual branch and an edge branch. The contextual branch captures remotely relevant information within the image to acquire global contextual information and enhance image components. The edge branch utilizes Sobel operators in two different directions to compute the image gradient, thereby detecting edges and enhancing the texture features of the image.

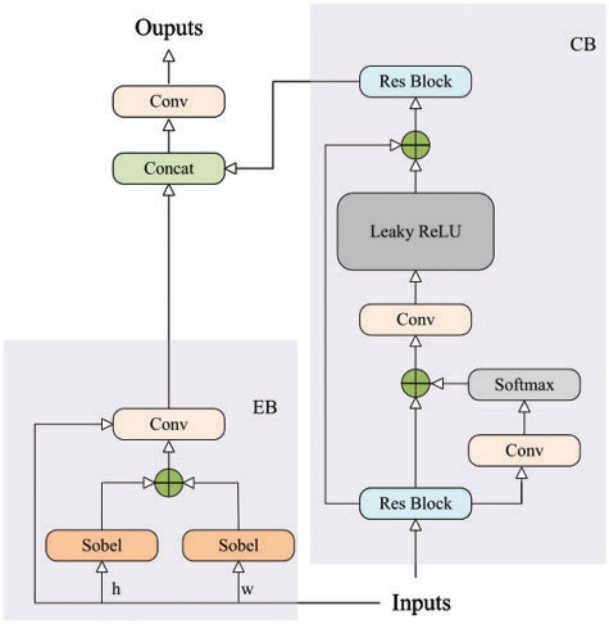


Figure 4: The detailed diagram of DPM

Contextual Branch: We use residual blocks to handle features before and after capturing long-range dependencies. Residual building blocks excel at relaying robust low-frequency details via direct shortcut pathways. The first residual block transforms the number of feature channels from 3 to 32, while the second residual block reverses this operation, Empirical evidence supports the notion that absorbing contextual global insights about a scene proves invaluable for elementary visual chores like bolstering image quality in limited light situations [35]. It is defined as:

$$CB(x) = x + \delta(f_1(\hat{x})) \tag{7}$$

where $\hat{x} = \sigma(F_2(x)) \cdot x$, F is the convolutional layer with kernel 3×3, δ here means Leaky ReLU activation, meanwhile, σ speaks to the softmax operation.

Edge Branch: The Sobel operator combines Gaussian filtering and differentiation to approximate edge detection by computing gradients. As a discrete operator, it detects edges by computing gradients. We apply the Sobel operator separately in the horizontal and vertical directions, we differentiate in the x-direction to obtain edges in the y-direction and differentiate in the y-direction to obtain edges in the x-direction. Edge information is re-extracted through convolutional filtering, and residual connections are utilized to facilitate information flow. This process can be formulated as:

$$EB(x) = F_3(sobel_h(x) + sobel_w(x)) + x \tag{8}$$

where $sobel_h$ and $sobel_w$ represent the Sobel operation in the vertical and horizontal directions, respectively.

3.2.3 Learnable Deep-Guided Filter

Guided filtering [36] is an operation focused on edge preservation and gradient fidelity. Its core idea is to enhance the prominence of target edges by utilizing boundary information of objects in the guiding image. This method improves downstream tasks like object detection or image segmentation by minimizing the saliency of regions outside target contours. It takes an original image and a guiding image as inputs, where the processed image maintains the overall appearance of the original but approximates the texture details of the guiding image. Essentially, the guiding image represents the gradient map of the original, highlighting high gradients at edges and low gradients in flat regions.

Contemporary advanced systems developed for visual problem solving increasingly employ guidance filters to render more polished outcomes, as seen [37,38]. The guided filter takes a low-resolution image I_L, its corresponding high-resolution image I_h, and a low-resolution output O_L as inputs, generating a high-resolution output O_h. Specifically, A_L and b_L are first optimized to minimize the reconstruction error between O_L and $\hat{O}_L$, where $\hat{O}_L$ follows a local linear model:

$$\hat{O}_L^i = A_L^k I_L^i + b_L^k, \forall i \in \varpi k \tag{9}$$

where (A_L^k, b_L^k) represents the linear coefficients, and ϖk is the square window of radius r centered at the k-th local square window on I_L. I_L^i is the i-th pixel in image ϖk, followed by upsampling A_L and b_L to obtain A_h and b_h. The high-resolution output O_h is ultimately generated by a linear transformation model, and specific inferences can be referred to [39].

$$O_h = A_h * I_h + b_h \tag{10}$$

To enhance segmentation quality, we introduced the Depth Guided Filter (DGF) during post-processing. DGF, proposed in reference [39], replaces mean filtering with dilated convolutions and employs convolutional blocks composed of pointwise convolutions. With learnable parameters, DGF adapts dynamically to specific needs, facilitated by terminal-point training. The architecture diagram of the Depth Guided Filter is shown in Fig. 5 and the algorithm flow is shown in Table 2. The transformation function $F(I)$ converts I_h and I_L into task-oriented maps G_h and G_L. $F(I)$ consists of an FCN block with two convolutional layers, an adaptive normalization layer, and a ReLU layer in between. The convolutional kernels have dimensions of 1×1, with 64 and 19 output channels, respectively.

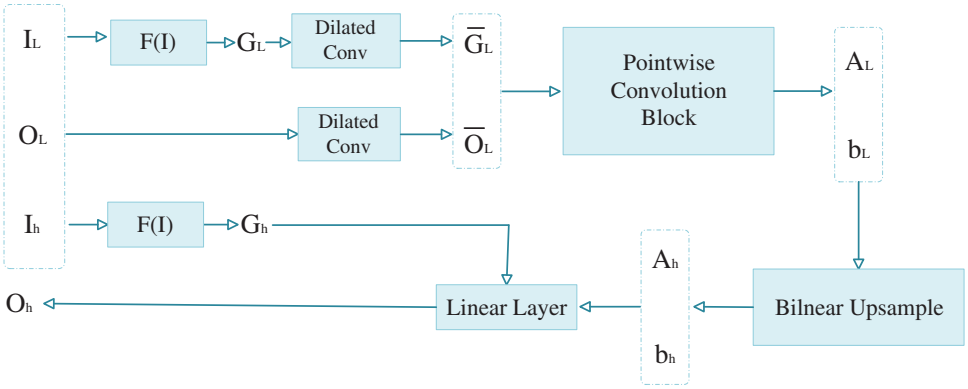


Figure 5: The computation graph of the DGF. Here, I_L represents the IAEN processed output, O_L represents the output of the segmentation network, and O_h represents the final output

Table 2: DGF algorithm structure

Algorithm: Learnable Deep Guided Filter
Input: segmentation network output (O_L)
$G_L = F(I_L)$
$G_h = F(I_h)$
$\overline{G}_L = fmean(G_L)$
$\overline{O}_L = fmean(O_L)$
$varG_L = corrG_L - meanG_L \times meanG_L$
$covG_LO_L = corrG_LO_L - meanG_L \times meanO_L$
$A_L = cov\overline{G_LO_L}/var\overline{G}_L$
$b_L = meanO_L - A_L * mean\overline{G}_L$
$A_h = A_L f \uparrow$
$b_h = b_L f \uparrow$
$O_h = A_h * G_h + b_h$

3.3 Probability Re-Weighting

In driving scene images, the pixel counts for different object categories are uneven, making it challenging for the network to learn features of small-sized objects during training. Building upon point [6], we employ a reweighting scheme to enhance the network's focus on small-sized objects. The reweighting formula is defined as follows:

$$w_i' = -\log(a_i) \tag{11}$$

where a_i expresses the percentage of marked pixels belonging to group i within any provided still. Conspicuously, smaller magnitudes of a_i imply heavier importance placed upon them. Using such weights helps the network focus on small-sized objects and effectively learn their features, giving greater attention to small objects and thereby improving the network' is performance in segmenting small object categories. The normalization of weights for each class is as follows. To ensure that the

value of w_i is positive, we preselect two parameters, std and avg. During training, we set $std = 0.05$ and $avg = 1.0$.

$$w_i = \frac{w_i' - \overline{w}}{\sigma(x)} std + avg \tag{12}$$

where $\overline{w}$ and $\sigma(x)$ represent the mean and standard deviation of w_i' respectively.

4 Experiments

4.1 Datasets and Evaluation Metrics

For all experiments, the evaluation metric is the average of the class intersection over union (mIOU), which quantifies the accuracy of the model by computing the intersection over union ratio of each class between the predicted segmentation results and the actual annotations. The calculation of mIOU follows the mathematical expression below:

$$mIOU = \frac{1}{n} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{13}$$

where N is the number of classes, TP_i is the true positive count for the i class, FP_i is the false positive count for the i class, and FN_i is the false negative count for the i class. The datasets used for training and evaluating the model are as follows:

Cityscapes [40]: Cityscapes is a dataset focused on urban street scenes during the day, serving as a benchmark for semantic segmentation tasks. The dataset holds pixel-granular tags covering 19 groups, providing 2975 images destined for teaching purposes, another 500 kept apart for ability confirmation, and finally 1525 earmarked for stringent practical runs.

Nightcity [4]: Nightcity is a large dataset of pixel-level annotated nighttime urban driving scenes, consisting of 2998 images for training and 1299 images for validation or testing. The annotated object categories are consistent with Cityscapes.

ACDC [5]: ACDC is a dataset with pixel-level annotations for adverse driving conditions, featuring 4006 high-quality images across four common scenarios: foggy, nighttime, rainy, and snowy conditions. For our supervised experiments, we used ACDC-night, which includes 400 training images, 106 validation images, and 500 testing images. Additionally, we evaluated the model's generalization using 1000 foggy scene images, with 800 for training and 200 for testing.

4.2 Experimental Settings

All experiments used pre-trained semantic segmentation models trained for 150,000 iterations on Cityscapes. We implemented our approach in PyTorch on a single Nvidia A6000 GPU. Following reference [6], we used Stochastic Gradient Descent (SGD) with a momentum of 0.9, weight decay of 5×10^{-4}, and an initial learning rate of 2.5×10^{-4} reduced by a factor of 0.9. The batch size was 4. Data augmentation included random cropping (0.5 to 1.0 scale, 512×512 dimensions) and random horizontal flips.

4.3 Comparison with State-of-the-Art Methods

Experiments on Cityscapes and Nightcity Datasets: To validate the effectiveness of the proposed method in image segmentation tasks, we conducted experiments on four different datasets, comparing our approach with state-of-the-art methods. Table 3 presents the quantitative results of existing

methods and our proposed method on the Nightcity dataset. Our method uses DeepLabV2, PSPNet, and RefineNet as baseline models, and compared to previously proposed networks, it demonstrated performance improvements of 3.23%, 4.21%, and 3.95% on these baselines, respectively. Compared to the current state-of-the-art method (Liu et al. [9]), our method achieved an improvement of 2.2%. Furthermore, our method also exhibited better test results on the daytime Cityscapes dataset. Fig. 6 illustrates visual comparisons between our approach, the baseline model PSPNet, and state-of-the-art methods [9]. As shown in the figures, our method excels in nighttime segmentation by capturing finer details that other models tend to overlook, such as traffic signs.

Table 3: Visual segmentation results of our method and baseline model on Nightcity test set, where "C" represents trained on Cityscapes. "C + N" represents trained on Cityscapes and Nightcity. "N_t" represents the Nightcity test and "C_v" represents the Cityscapes validation set

Methods	mIOU (%) on N_t		mIOU (%) on C_v	
	C	C + N	C	C + N
DeepLabV2 [17]	18.20	46.39	66.37	66.65
Liu et al. [9]	–	48.24	–	66.57
Ours	–	49.67	–	64.98
PSPNet [19]	20.65	47.29	63.94	65.87
Liu et al. [9]	–	49.73	–	66.59
Ours	–	52.11	–	66.61
RefineNet [20]	22.92	48.70	65.85	65.68
Liu et al. [9]	–	51.21	–	67.15
Ours	–	**52.65**	–	**67.34**

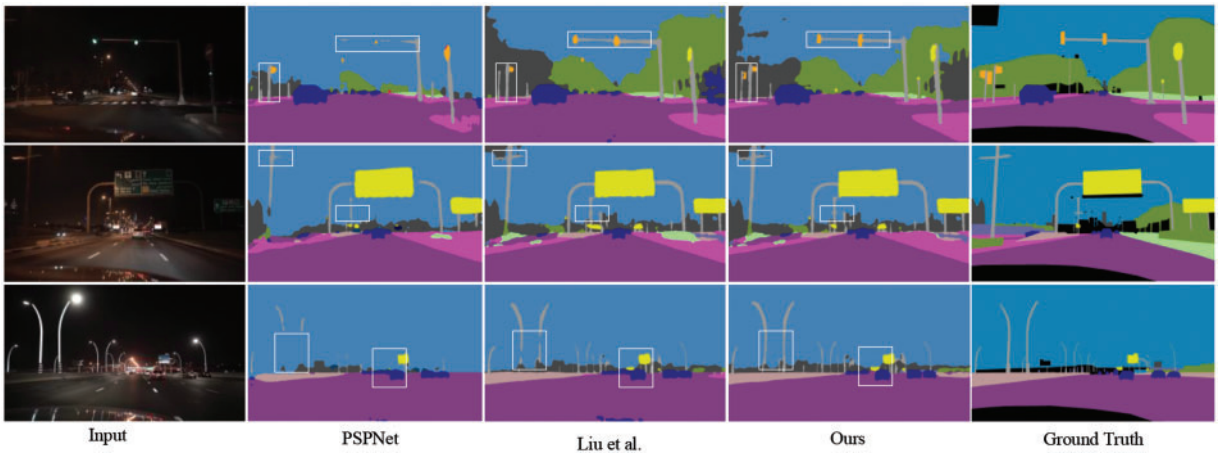


Figure 6: The visual segmentation results of our method and existing models on the Nightcity dataset

Experiments on Cityscapes and ACDC-Night Datasets: We evaluated the effectiveness of our proposed method on a mixed dataset combining Cityscapes and ACDC-night, as shown in Table 4. On the ACDC-night dataset, our method continues to demonstrate strong performance. Compared to prior work [9], our method shows improvements across all three benchmark models. Notably, our

method achieves the best segmentation performance on the RefineNet benchmark model. As depicted in the Fig. 7, the image adaptive module can effectively distinguish objects of interest from the image, particularly small objects in areas with mixed categories under low light conditions.

Table 4: Visual segmentation results of our method and baseline model on ACDC-night test set, where "C" represents trained on Cityscapes. "C + AN" represents trained on Cityscapes and ACDC- night. "AN_t" represents ACDC-night test, and "C_v" represents Cityscape validation test

Methods	mIOU (%) on AN_t		mIOU (%) on C_v	
	C	C + AN	C	C + AN
DeepLabV2 [17]	30.06	53.31	66.37	64.97
Liu et al. [9]	–	55.78	–	66.55
Ours	–	57.98	–	64.98
PSPNet [19]	26.62	56.69	63.94	65.18
Liu et al. [9]	–	58.42	–	66.75
Ours	–	59.66	–	66.65
RefineNet [20]	29.05	57.69	65.85	63.19
Liu et al. [9]	–	60.06	–	66.19
Ours	–	**61.26**	–	**66.94**

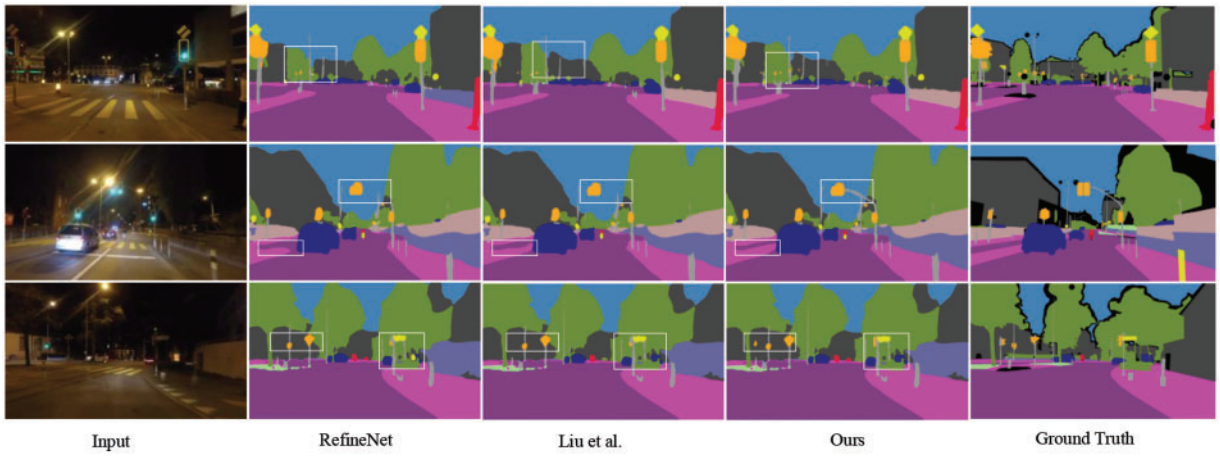


Figure 7: The visual segmentation results of our method and existing models on the ACDC-night dataset

Experiments on Cityscapes and ACDC-Fog Datasets: To validate the effectiveness and scalability of our method, as well as the autonomous selectivity of the gating mechanism, we conducted experiments on a mixed dataset combining Cityscapes and ACDC-fog. Table 5 presents the quantitative results of our method on ACDC-fog. Based on the results, our method exhibits higher accuracy in handling foggy weather images. This is mainly attributed to the dehazing filter in the image adaptive network. Fig. 8 illustrates different parameter predictions in low-light and foggy weather images, indicating that our image adaptive network adjusts adaptively based on different input images.

Table 5: Visual segmentation results of our method and baseline model on ACDC-fog test set, where "C" represents trained on Cityscapes. "C + AF" represents trained on Cityscapes and ACDC-fog, "AF_t" represents ACDC-fog test, and "C_v" represents Cityscape validation test

Methods	mIOU(%) on AF_t		mIOU(%) on C_v	
	C	C + AF	C	C + AF
DeepLabV2 [17]	33.47	52.19	66.37	64.21
Ours	–	57.97	–	65.63
PSPNet [19]	39.54	47.29	63.94	66.8
Ours	–	69.74	–	**67.55**
RefineNet [20]	46.39	66.65	65.85	65.76
Ours	–	**70.9**	–	65.92

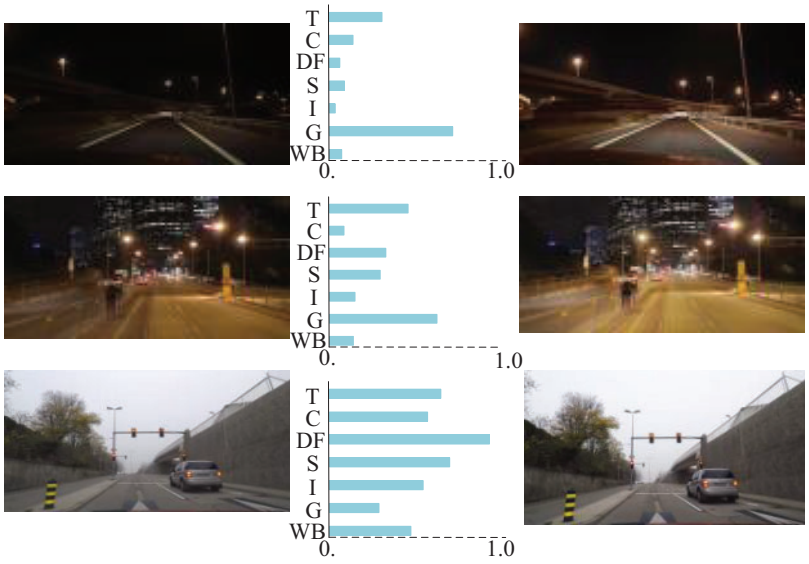


Figure 8: The bar chart displays the emission gates of different input images. Edip applies the optimal enhancement method learned based on the input image features

4.4 Ablation Study

To validate the effectiveness of each module in our proposed framework, including DGF, DPM, and Edip, we conducted ablation experiments with various configurations. All experiments were conducted on the combined Cityscapes and Nightcity datasets for training. All ablation experiments are conducted using PSPNet as the baseline model. The experimental results are shown in Table 6. It can be observed that, each module we designed positively impacts nighttime image segmentation.

Table 6: Dissolution experiments of our method on the Nightcity dataset, where "N_t" represents the Nightcity test and "C_v" represents the Cityscapes validation set

Method	mIOU (%) on N_t	mIOU (%) on C_v
PSPNet [19]	47.29	65.87
Edip (SE)	50.02	65.93
Edip (Max)	49.65	66.72
Edip (Line)	50.72	67.42
w/o DPM	50.26	65.94
w/o DGF	51.37	66.62
Ours	**52.65**	**67.34**

Single Best *vs*. Weighted Combination: The gating mechanism we proposed aids in combining image processing operations by using relative weights see the Eq. (1). To demonstrate its effectiveness, we removed this mechanism and utilized a single optimal image processing operation based on a single gate value (referred to as "Edip (Max)" in the Table 6). For the Nightcity dataset, segmentation performance decreased by 3% in the absence of the proposed gating mechanism. This study demonstrates the crucial importance of integrating multiple image enhancement techniques because no single method alone can adequately handle the demands of processing complex nighttime images.

Single-Layer Embedding *vs*. Multi-Layer Embedding: To examine the effectiveness of employing embeddings at different levels to facilitate Edip's access to multiple feature scales, we compared the results with a single-level embedding, denoted as "Edip (SE)" in Table 6. In comparison to using multi-scale embeddings, segmentation performance decreased by 2.63%.

Simultaneous *vs*. Production Line: Additionally, we adaptively process input images in a pipeline manner, referred to as "Edip (Line)" in Table 6. For the Nightcity dataset, employing a sequential pipeline results in a 1.93% decrease in segmentation performance. Fig. 9 depicts the simultaneous extraction of multiple image processing parameters. This study emphasizes the importance of concurrently extracting these parameters to avoid inaccuracies in revealing crucial image details, particularly for nighttime image segmentation.

With and without DPM: The proposed detail refinement module showed a decrease of 2.39% in segmentation performance for the Nightcity dataset when DPM was not included. As shown in Fig. 10, it demonstrates the visual results with and without the DPM module, highlighting the advantageous impact of introducing DPM for segmenting small objects.

With and without DGF: The proposed Depth Guided Filter (DGF) used for post-processing further enhances performance. For the Nightcity dataset, segmentation performance decreased by 1.28% when DGF was not included. Additionally, Fig. 11 illustrates the visual results with and without DGF. It can be observed that introducing the Depth Guided Filter is advantageous for handling segmentation boundaries.

Real-Time Performance: In our research, we conducted real-time speed performance comparisons on a single Nvidia A6000 GPU, the experimental results are shown in Table 7, benchmarking our network against other techniques. The results indicate that while we have improved accuracy, there is only approximately a 6 FPS difference compared to the latest model. We plan to further optimize the model structure to achieve a dual enhancement in both speed and accuracy.

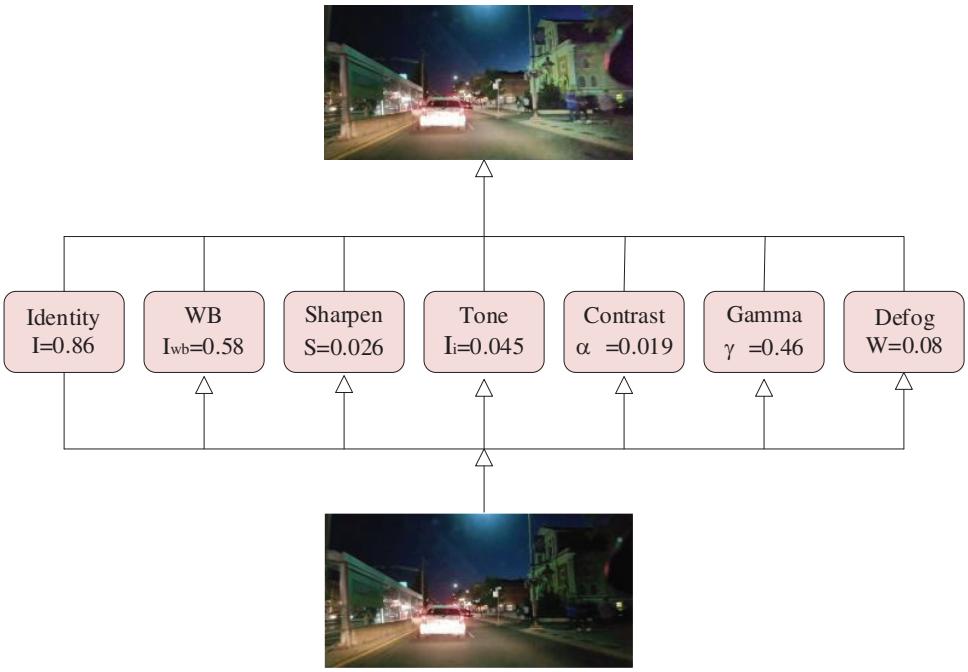
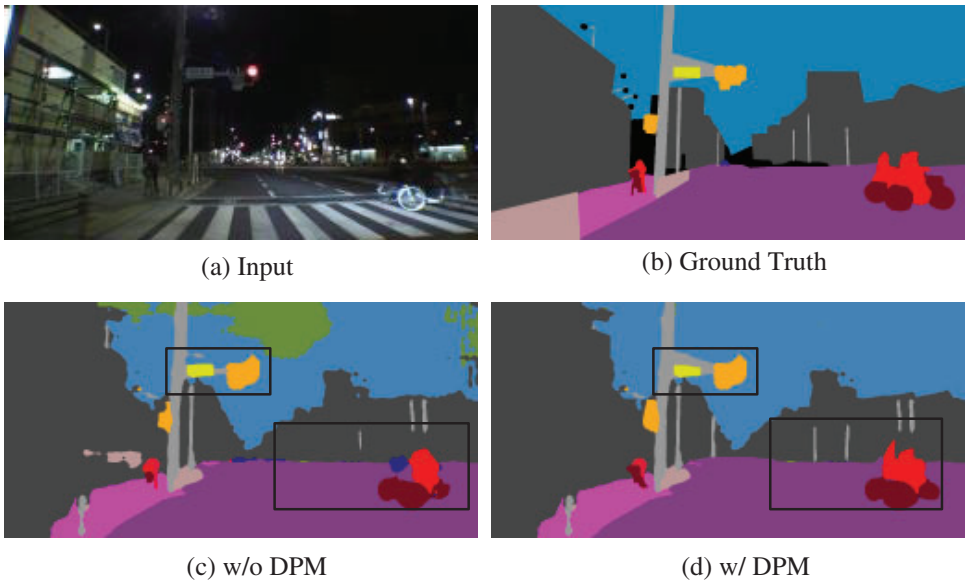


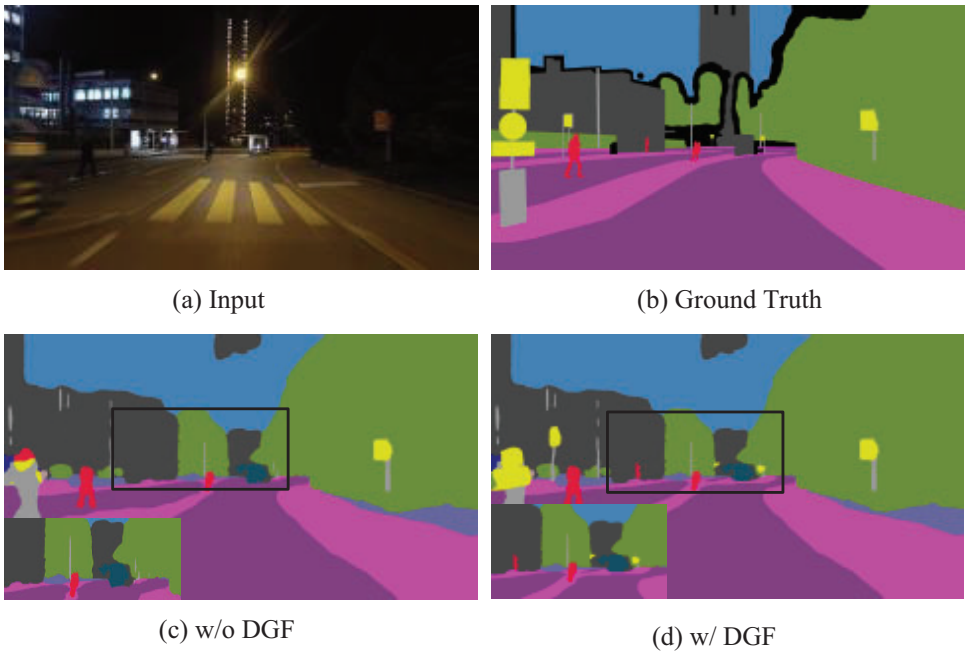
Figure 9: An example of simultaneously predicting multiple parameters



(a) Input (b) Ground Truth

(c) w/o DPM (d) w/ DPM

Figure 10: The visualization results of our method on the DPM module with and without on the Nightcity test set samples



(a) Input (b) Ground Truth

(c) w/o DGF (d) w/ DGF

Figure 11: The visualization results of our method on the DGF module with and without on the Nightcity test set samples

Table 7: Real-time performance on Nvidia A6000 GPU

Method	FPS
PSPNet [19]	34.72 ± 1.5
Liu et al. [9]	30.61 ± 3.1
Ours	25.26 ± 2.7

5 Conclusion

This paper presents a novel nighttime image segmentation framework comprising an Image Adaptive Enhancement Network (IAEN) integrated into the segmentation network's head and a Depth Guided Filter (DGF) in the network's tail. IAEN includes a parameter predictor (Edip), multiple image processing modules (Mdif), and a Detail Processing Module (DPM). The framework employs different levels of embedding to improve prediction accuracy by providing Edip access to multiple feature scales. A gating mechanism in Edip enables the concurrent operation of various image processing modules, with their weights combined via predicted gate weights to produce the final enhanced image. DGF further enhances segmentation accuracy. The end-to-end training of this framework demonstrates its versatility across autonomous driving scenarios, contributing to enhanced safety features such as collision avoidance and nighttime pedestrian recognition, and accelerating the adoption of autonomous driving technology.

Acknowledgement: We want to express our gratitude to the editors and every reviewer for their hard work and professional advice throughout the process. Their careful review and valuable feedback have made significant contributions to the improvement of this paper.

Funding Statement: This work is supported in part by The National Natural Science Foundation of China (Grant Number 61971078), which provided domain expertise and computational power that greatly assisted the activity. This work was financially supported by Chongqing Municipal Education Commission Grants for-Major Science and Technology Project (Grant Number gzlcx20243175).

Author Contributions: The authors confirm contribution to the paper as follows: Xiaoli Yuan: Designing methodologies, crafting network modules, coding, and thesis composition. Jianxun Zhang: Guides the work and analyzes the theoretical nature of the module. Xuejie Wang: Dataset analysis, inference result uploading. Zhuhong Chu: Review, supervision, and collect recent papers. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Date openly available in a public repository, code is available at https://github.com/lucky-yxl/ED-Ged, accessed on 25 April 2024.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] P. S. Chib and P. Singh, "Recent advancements in end-to-end autonomous driving using deep learning: A survey," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 103–118, Jan. 2024. doi: 10.1109/TIV.2023.3318070.

[2] H. Li, Y. Iwamoto, X. Han, A. Furukawa, S. Kanasaki and Y. W. Chen, "An efficient and accurate 3D multiple-contextual semantic segmentation network for medical volumetric images," in *2021 43rd Annual Int. Conf. IEEE Eng. Med. & Bio. Soc. (EMBC)*, Mexico, 2021, pp. 3309–3312. doi: 10.1109/EMBC46164.2021.9629671.

[3] L. Han, T. Zheng, Y. Zhu, L. Xu, and L. Fang, "Live semantic 3D perception for immersive augmented reality," *IEEE Trans. Vis. Comput. Graph*, vol. 26, no. 5, pp. 2012–2022, May 2020. doi: 10.1109/TVCG.2020.2973477.

[4] X. Tan, K. Xu, Y. Cao, Y. Zhang, L. Ma and R. W. H. Lau, "Night-time scene parsing with a large real dataset," *IEEE Trans. Image Process.*, vol. 30, pp. 9085–9098, Nov. 2021. doi: 10.1109/TIP.2021.3122004.

[5] C. Sakaridis, D. Dai, and L. V. Gool, "ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *2021 IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 10745–10755. doi: 10.1109/ICCV48922.2021.01059.

[6] X. Wu, Z. Wu, H. Guo, L. Ju, and S. Wang, "DANNet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation," in *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, 2021, pp. 15764–15773. doi: 10.1109/CVPR46437.2021.01551.

[7] C. Sakaridis, D. Dai, and L. V. Gool, "Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3139–3153, 1 Jun. 2022. doi: 10.1109/TPAMI.2020.3045882.

[8] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang and F. Wen, "Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation," in *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, 2021, pp. 12409–12419. doi: 10.1109/CVPR46437.2021.01223.

[9] W. Liu, W. Li, J. Zhu, M. Cui, X. Xie and L. Zhang, "Improving nighttime driving-scene segmentation via dual image-adaptive learnable filters," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 10, pp. 5855–5867, Oct. 2023. doi: 10.1109/TCSVT.2023.3260240.

[10] X. Chen *et al.*, "Dual adversarial attention mechanism for unsupervised domain adaptive medical image segmentation," *IEEE Trans. Med. Imaging*, vol. 41, no. 11, pp. 3445–3453, Nov. 2022. doi: 10.1109/TMI.2022.3186698.
[11] G. Monica, C. Nicola, and E. Ugo, "Egocentric upper limb segmentation in unconstrained real-life scenarios," *Virtual Real.*, vol. 27, no. 4, pp. 3421–3433, Oct. 2023. doi: 10.1007/s10055-022-00725-4.
[12] Y. Sun, Z. Ma, M. Zhou, and Z. Cao, "A topological semantic mapping method based on text-based unsupervised image segmentation for assistive indoor navigation," *IEEE Trans. Instrum. Meas.*, vol. 72, no. 2531513, pp. 1–13, 2023. doi: 10.1109/TIM.2023.3326167.
[13] Y. Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images," in *Proc. Eighth IEEE Int. Conf. Comput. Vis. ICCV 2001*, Vancouver, BC, Canada, 2001, vol. 1, pp. 105–112. doi: 10.1109/ICCV.2001.937505.
[14] N. Y. An and C. M. Pun, "Iterated graph cut integrating texture characterization for interactive image segmentation," in *2013 10th Int. Conf. Comput. Graph., Imag. Visual.*, Los Alamitos, CA, USA, 2013, pp. 79–83. doi: 10.1109/CGIV.2013.34.
[15] J. Long, S. Evan, and D. Trevor, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Jan. 2016. doi: 10.1109/TPAMI.2016.2572683.
[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," presented at Int. Conf. Medical Image Comput. Comput.-Assist. Interven. Springer International Publishing, Munich, Germany, Nov. 2015, pp. 234–241.
[17] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018. doi: 10.1109/TPAMI.2017.2699184.
[18] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 801–818.
[19] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 6230–6239. doi: 10.1109/CVPR.2017.660.
[20] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 5168–5177. doi: 10.1109/CVPR.2017.549.
[21] W. B. Zou, R. J. Long, Y. H. Zhang, M. X. Liao, Z. Zhou and S. S. Tian, "Dual geometric perception for cross-domain road segmentation," *Technol. Appl.*, vol. 76, no. 1, pp. 108–126, Jan. 2023. doi: 10.1016/j.displa.2022.102332.
[22] N. u. Khan, K. V. Arya, and M. Pattanaik, "An efficient image noise removal and enhancement method," in *2010 IEEE Int. Conf. Syst., Man and Cybernet.*, Istanbul, Turkey, 2010, pp. 3735–3740. doi: 10.1109/ICSMC.2010.5641838.
[23] Z. Yu and C. Bajaj, "A fast and adaptive method for image contrast enhancement," in *2004 Int. Conf. Image Process.*, Singapore, 2004, vol. 2, pp. 1001–1004. doi: 10.1109/ICIP.2004.1419470.
[24] J. Hai *et al.*, "R2RNet: Low-light image enhancement via real-low to real-normal network," *J. Vis. Commun. Image Rep.*, vol. 90, no. 15, pp. 103712, 2022. doi: 10.1016/j.jvcir.2022.103712.
[25] Z. Fu, Y. Yang, X. Tu, Y. Huang, X. Ding and K. K. Ma, "Learning a simple low-light image enhancer from paired low-light instances," in *2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, 2023, pp. 22252–22261. doi: 10.1109/CVPR52729.2023.02131.
[26] X. K. Liu, W. H. Ma, X. R. Ma, and J. Wang, "LAE-Net: A locally adaptive embedding network for low-light image enhancement," *Pattern Recognit.*, vol. 133, no. 4, pp. 109039–109050, 2023. doi: 10.1016/j.patcog.2022.109039.
[27] W. Liu, G. Yu, S. Guo, J. Zhu, and L. Zhang, "Image-adaptive YOLO for object detection in adverse weather conditions," *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 2, pp. 1792–1800, 2022. doi: 10.1609/aaai.v36i2.20072.

[28] R. Xia, C. Zhao, M. Zheng, Z. Wu, Q. Sun and Y. Tang, "CMDA: Cross-modality domain adaptation for nighttime semantic segmentation," in *2023 IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, 2023, pp. 21515–21524. doi: 10.1109/ICCV51070.2023.01972.

[29] Y. Zhang, P. David, H. Foroosh, and B. Gong, "A curriculum domain adaptation approach to the semantic segmentation of urban scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 1823–1841, 1 Aug. 2020. doi: 10.1109/TPAMI.2019.2903401.

[30] J. Vertens, J. Zürn, and W. Burgard, "HeatNet: Bridging the day-night domain gap in semantic segmentation with thermal images," in *2020 IEEE/RSJ Int. Conf. Intell. Robots and Syst. (IROS)*, Las Vegas, NV, USA, 2020, pp. 8461–8468. doi: 10.1109/IROS45743.2020.9341192.

[31] H. Wang *et al.*, "SFNet-N: An improved SFNet algorithm for semantic segmentation of low-light autonomous driving road scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 21405–21417, Nov. 2022. doi: 10.1109/TITS.2022.3177615.

[32] X. T. Yang, J. Y. Han, and C. Z. Liu, "A semantic segmentation scheme for night driving improved by irregular convolution," *Front Neurorobot*, vol. 17, pp. 214, Jun. 2023. doi: 10.3389/fnbot.2023.1189033.

[33] G. Yang, Z. Zhun, H. Tang, M. L. Ding, N. Sebe and E. Ricci, "Bi-Mix: Bidirectional mixing for domain adaptive nighttime semantic segmentation," arXiv preprint arXiv:2111.10339, 2021, pp. 163–174.

[34] S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," *Int. J. Comput. Vis.*, vol. 48, no. 3, pp. 233–254, Jul. 2002. doi: 10.1023/A:1016328200723.

[35] X. C. Yin, Z. D. Yu, Z. T. Fei, W. J. Lv, and X. Gao, "PE-YOLO: Pyramid enhancement network for dark object detection," in *Artificial Neural Networks and Machine Learning–ICANN 2023*, vol. 14260. Cham: Springer, Sep. 2023. doi: 10.1007/978-3-031-44195-0_14.

[36] K. M. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Oct. 2012. doi: 10.1109/TPAMI.2012.213.

[37] C. Chen, C. Zhang, C. Li, and J. Hong, "Assembly monitoring using semantic segmentation network based on multiscale feature maps and trainable guided filter," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022. doi: 10.1109/TIM.2022.3204322.

[38] X. Zhang, W. Zhao, W. Zhang, J. Peng, and J. Fan, "Guided filter network for semantic image segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 2695–2709, 2022. doi: 10.1109/TIP.2022.3160399.

[39] H. Wu, S. Zheng, J. Zhang, and K. Huang, "Fast end-to-end trainable guided filter," in *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 1838–1847. doi: 10.1109/CVPR.2018.00197.

[40] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 3213–3223. doi: 10.1109/CVPR.2016.350.