2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

End-to-end Semantic Segmentation Network for Low-Light Scenes

Hongmin Mu, Gang Zhang, MengChu Zhou, *Fellow, IEEE* and Zhengcai Cao, *Senior Member, IEEE*

***Abstract*— In the fields of robotic perception and computer vision, achieving accurate semantic segmentation of low-light or nighttime scenes is challenging. This is primarily due to the limited visibility of objects and the reduced texture and color contrasts among them. To address the issue of limited visibility, we propose a hierarchical gated convolution unit, which simultaneously expands the receptive field and restores edge texture. To address the issue of reduced texture among objects, we propose a dual closed-loop bipartite matching algorithm to establish a total loss function consisting of the unsupervised illumination enhancement loss and supervised intersection-over-union loss, thus enabling the joint minimization of both losses via the Hungarian algorithm. We thus achieve end-to-end training for a semantic segmentation network especially suitable for handling low-light scenes. Experimental results demonstrate that the proposed network surpasses existing methods on the Cityscapes dataset and notably outperforms state-of-the-art methods on both Dark Zurich and Nighttime Driving datasets.**

I. Introduction

AIMING to label each pixel of a given image to an object category, semantic segmentation is a fundamental computer vision task and benefits many applications such as robot perception [1], autonomous driving [2], and medical imaging [3]. Although deep learning techniques have made significant advancements in the performance of semantic segmentation for images captured during the daytime under favorable lighting conditions, limitation significantly undermines its performance for low-light scenes [4]. To address this issue, many low-light image enhancement networks have been developed [5]–[7], but each requires independent training before integration into semantic segmentation. This separation prevents these networks from fully optimizing their parameters for the specific needs of the subsequent segmentation tasks. Furthermore, the complex standalone training process constrains the broad application of segmentation algorithms for low-light scenes.

This work is supported in part by the National Natural Science Foundation of China under Grant (92148202, 52175002, 52105005), FDCT under Grant No. 0047/2021/A1, and the Beijing Natural Science Foundation (L223019, 3242011). (Corresponding authors: Z. Cao and M. Zhou.)

Hongmin Mu and Gang Zhang are with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, 100029, China. (e-mail: 2021200806@buct.edu.cn, 2022210504@buct.edu.cn)

MengChu Zhou is with the Department of Macao Institute of Systems Engineering, Macau University of Science and Technology, Macao 999078, China, and also with the Helen and John C. Hartmann Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102, USA (e-mail: mengchu@gmail.com).

Zhengcai Cao is with the State Key Laboratory of Robotics and Systems, Harbin Institute of Technology, Harbin, 150080, China (e-mail: caozc@hit.edu.cn)

Significant progress has been made in the field of semantic segmentation of low-light scenes. Dai et al. [8] introduced an intermediate twilight domain to adapt semantic models trained on daytime scenes to nighttime ones progressively. Sakaridis et al. [9] extended the approach in [8] to a guided curriculum adaptation framework, utilizing both stylized synthetic images and unlabeled real images to leverage cross-time-of-day scene correspondence. However, these gradual adaptation approaches typically require training multiple semantic segmentation models, such as three models in [10] for three different domains, which is inefficient. Subsequent work along this direction [11]–[13] also trains additional image transfer models, but the performance of semantic segmentation heavily relies on the pre-trained image transfer model in the preprocessing stage.

In the area of low-light image segmentation, we identify two predominant challenges: 1) Targets appear with limited visibility, and 2) The color contrast at edge textures among distinct targets is subdued, hindering precise edge delineation. Our proposed network aims fundamentally crafted to address them.

To the best of our knowledge, there is no end-to-end semantic segmentation network designed for low-light conditions existed. The primary challenge stems from the inability of the existing methods to optimize the unsupervised illumination enhancement (IE) loss and supervised intersection-over-union (IoU) loss jointly. This issue is due to potential conflicts between the optimization objectives of these two losses. To address it well, we propose a Dual Closed-loop Bipartite Matching (DCBM) algorithm, which establishes a comprehensive loss criterion. Joint optimization of these two losses is used to perform the Hungarian algorithm [14]. Furthermore, we propose a Hierarchical Gated Convolution (HGC) unit. It not only expands the effective receptive field over the original network while maintaining an almost unchanged parameter count, but also leverages the potent edge feature extraction capability of the gated convolution, helping mitigate edge loss in low-light images.

In this work, we propose an End-to-end Semantic Segmentation Network for Low-light scenes (ESSNL), which learns to enhance images in low-light conditions for semantic segmentation in an end-to-end manner. This work intends to make the following new contributions:

1) It proposes an HGC unit that can extract more edge texture information than the original network and expands its receptive field on a fine-grained level, while maintaining an almost unchanged parameter count.

2) It proposes a DCBM algorithm that minimises both unsupervised IE loss and supervised IoU loss, thus achieving

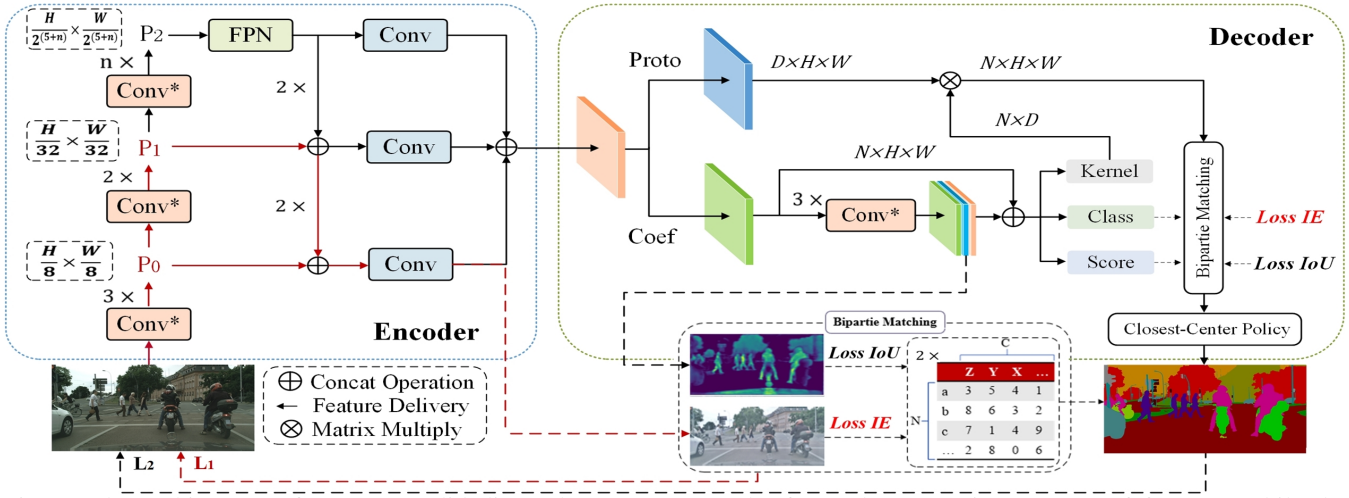


Fig. 1: The architecture of ESSNL. Solid directed arcs represent the forward propagation process of features, while the dashed ones below L_1 and L_2 indicate the backward propagation process of training losses. ESSNL achieves end-to-end training through the inner loop (indicated by the red directed arcs) and the outer loop, to be detailed in Sec II.3. The details of each stage (P0, P1, P2, etc.) are depicted as shown in Fig. 3, with the upper left corner indicating its resolution.

end-to-end training for the semantic segmentation network under low-light conditions.

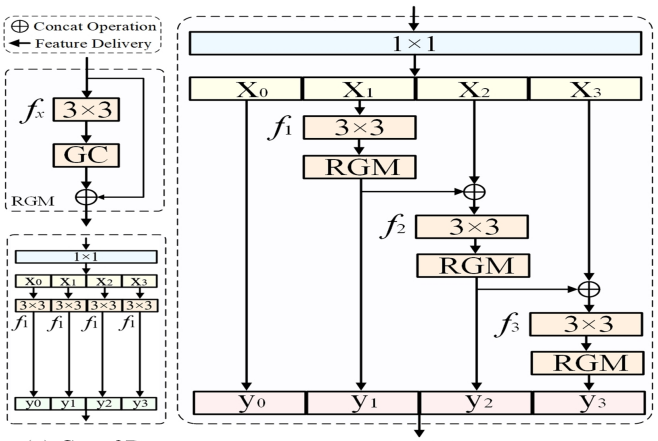
3) It designs the decoder of ESSNL based on mask-to-mask predictions rather than pixel-to-pixel ones, thus reducing erroneously predicted pixels inside each classified area. Experimental results on the Dark Zurich, Nighttime Driving and Cityscapes Datasets demonstrate the effectiveness of ESSNL.

II. Proposed Methods

The architecture of the proposed ESSNL is shown in Fig. 1. To enhance the class consistency inside each segmented area, we design the decoding part to make mask-based predictions instead of pixel-to-pixel ones. We employ a two-stream network structure [15]. To address mask overlap, we adopt the Closest-Center Policy [16]. Two bipartite graphs are utilized for loss calculation: one for unsupervised training of the light-enhanced shallow network and the other for supervised training to improve the segmentation accuracy of ESSNL. These graphs are interconnected through a dual closed-loop training strategy to facilitate co-optimization.

2.1 HGC Unit

Since gated convolution [17] has demonstrated its ability to extract edge features, we introduce it into the convolution unit in the form of hierarchical cascades, thus improving the ability to distinguish the edges of objects in low-light images. To endow the acquired features with multi-scaled receptive fields, we construct HGC as shown in Fig. 2(b). HGC first divides a feature map input into s subsets, represented by x_i, $i \in \{0, 1, \cdots, s-1\}$ and s is the number of grouping dimensions. Each subset of the feature map retains the same resolution *w.r.t* the input feature map, while the number of channels becomes $1/s$ *w.r.t* it. Every x_i except x_0 is input through a convolution layer with kernel size of 3×3 and a residual gated model (RGM) with the same number of channels w.r.t it, represented by $f_i()$. y_i is the output feature



(a) Conv2D (b) HGC

Fig. 2: The comparison between conventional 2D convolution (Conv2D) unit and HGC unit with the grouping dimension $s = 3$.

map of $f_i()$. The output is $\sum_{i=1}^{s} y_i$, where the addition refers to the concat operation of the obtained feature map. We have:

$$y_i = \begin{cases} x_i & i = 0 \\ f_i(x_i) & i = 1 \\ f_i(x_i + y_{i-1}) & 1 < i \leq s \end{cases} \quad (1)$$

To compare the size of receptive fields (SRF) of Conv2D and HGC, we assume that both convolutions operate on the feature map L_i of kernel size f_k and stride S_i. Considering that only one kernel size is used, *i.e.*, K_s in HGC, we have the ratio of SRF between Conv2D and HGC:

$$\frac{R_1}{R_k} = \frac{1 + (K_s - 1)S_1}{1 + (K_s - 1)\sum_{i=0}^{k}(\prod_{j=0}^{i} S_j)} \quad (2)$$

where K_s is set as 3 to obtain the largest SRF with the same number of parameters in Conv2D operation [18]. The SRF

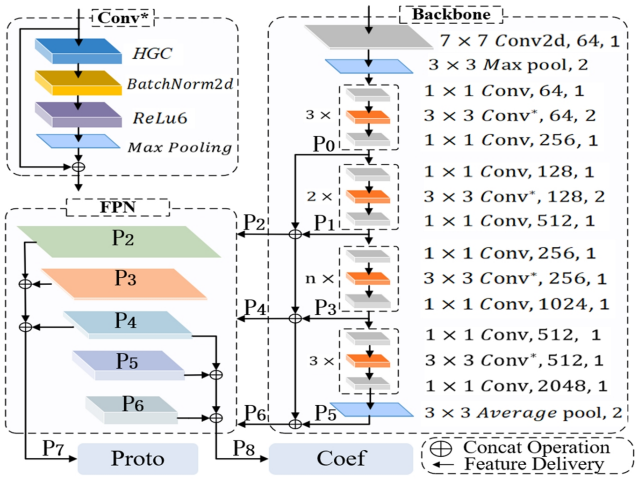


Fig. 3: The Encoder architecture of ESSNL. The (3×3, *conv*, 64, 2) represents HGC operation with kernel size of 3×3, kernels of 64, and stride of 2. P_i represents the feature map obtained by convolution operations. Proto and Coef [19] are feature maps shown in Fig. 2

of y_0, y_1, y_2, and y_3 in Fig. 2 (b) are 1, R_1, R_2, and R_3, respectively. It demonstrates that HGC endows an obtained feature with multi-scale receptive fields (1, R_1, R_2 and R_3), and its maximum receptive field is about 2 to 6 times larger than the single receptive field of Conv2D.

HGC units are utilized in the Encoder of ESSNL, as shown in Fig. 3. Among the four groups of residual modules, only the number n of the third group changes, referring to the architectures of ResNet50 to 101 [20].

2.2 Bipartite Matching for Loss Formulation

To enhance the brightness, we adopt the loss functions that are well designed by Zero-DCE++ [5]:

$$L^e = w_1 L_1^e + w_2 L_2^e + w_3 L_3^e, \tag{3}$$

where L_1-L_3 are the exposure control loss, color constancy loss, and illumination smoothness loss, respectively. We use the same weights as [5]'s, i.e., $w_1 = 1, w_2 = 0.5, w_3 = 20$.

The proposed ESSNL generates two fixed-size arrays of predictions: one encompassing light enhancement images (as illustrated before Loss IE in Fig. 1), and the other for semantic masks (as illustrated before Loss IoU in Fig. 1). Instead of grappling with the complexity of manually aligning ground truth entities using handcrafted rules, we employ bipartite matching to enable an efficient end-to-end training approach.

To tap the end-to-end training, we formulate the label assignment problem as a biparty graph matching one. Then we use the Hungarian algorithm [14] to find the minimum value of the bipartite graph as the loss value:

$$\mathcal{C}(i,k) = p_{i,c_k}^{1-\alpha} . \text{DICE}(m_i, t_k)^{\alpha}, \tag{4}$$

$$L^h = \sum_{i=1}^{N} \left[-\log(p_{\hat{\sigma}(i)}(c_i)) + C(i,k) \right] \tag{5}$$

where α is a hyper-parameter that balances the influences of classification and segmentation. c_k represents the category label for the k-th ground-truth object, and p_{i,c_k} indicates the probability of the i-th prediction belonging to category c_k. m_i and t_k are the masks of the i-th prediction and the k-th ground-truth object, respectively. N is the number of matched pairs, and $p_{\hat{\sigma}(i)}(c_i)$ is the predicted probability for a class.

2.3 Dual Closed-loop for Loss Optimization

Based on our experimental results, creating the total loss as a weighted summation of two individual losses leads to a non-convergence issue during training. This can be attributed to the potential conflicting optimization objectives of the two individual losses. To address this challenge, rather than merely weighting and summing the losses, we propose a dual-loop training strategy.

1) **Inner loop** focuses on unsupervised training, thereby enhancing the illumination of activation maps. The loss of this loop (L_1) reflects how well the encoder's shallow IE network is performing this task.

2) **Outer loop** Takes charge of the supervised training to enhance semantic segmentation precision. The loss of this loop (L_2) indicates the accuracy of semantic segmentation of ESSNL.

Our idea is to fine-tune the inner loop and add its loss value L_1 to the outer loop for training. We multiply the loss of the inner loop by a weight β_1 for back propagation, denoted as $L()$. In the forward propagation of the outer loop, denoted as $F()$, we already incorporates the inner loop's values. We further multiply its loss by β_2 for back propagation. Finally, we add this to the inner loop's loss, which is weighted by β_3, *i.e.*:

$$\begin{cases} L_2(x_k) = F[L_2(x_{k-1}) + \beta_1 L_1(x_{k-1})], \\ L_1(x_k) = F[L_1(x_{k-1}) + \beta_2 L_2(x_{k-1})], \\ L_2^{\dagger}(x_1) = F[L_2(x_0) + \beta_3 L_1(x_0)], \\ L_1^{\dagger}(x_1) = \beta_3 L_1(x_0), L_2(x_0) = L^0, k \geq 2 \end{cases} \tag{3}$$

where $L_2^{\dagger}$ and $L_1^{\dagger}$ correspond to the initial values of L_2 and L_0, respectively. During the initial training phase, ESSNL conducts a single outer loop training to derive the initial loss L^0. L_2 is the initial value of $L_2^{\dagger}$ at next forward propagation.

This dual-loop approach ensures that both unsupervised training for image illumination-enhanced and supervised training for semantic segmentation accuracy are well integrated within ESSNL, thereby ensuring the convergence of the overall training process for effective end-to-end training.

III. Experimental Results

In this section, we evaluate the performance of the introduced ESSNL across three public datasets and present results from ablation studies.

3.1 Datasets and Evaluation Metrics

For all experiments, we use the mean of category-wise intersection-over-union (mIoU) as the evaluation metric. The

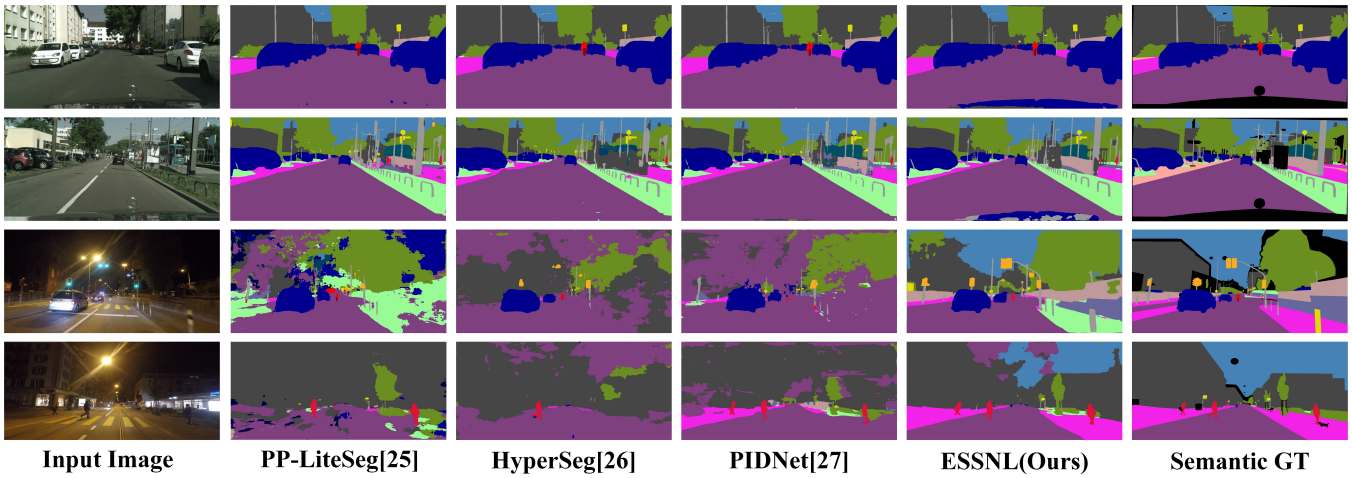


Fig. 4: Visualization comparison of our ESSNL with existing SOTA real-time methods on four samples from Cityscapes-val (the first 2 rows) and Dark-Zurich-val (the last 2 rows). Note that these networks are only trained with Cityscapes dataset.

TABLE I: Performance on Cityscapes val-test and test-set, where τ_i represents the number of processed frames per second.

Method	**mIoU**		τ_i	GPU	*Resoution*	*GFLOPs*	*Params*
	Val	Test					
SwiftNetRN-18 [21]	75.5	75.4	39.9	GTX 1080Ti	2048×1024	104	11.8M
SwiftNetRN-18-ens [21]	-	76.5	18.4	GTX 1080Ti	2048×1024	218	24.7M
CABiNet [22]	76.6	75.9	76.5	RTX 2080Ti	2048×1024	12	2.64M
BiSeNet (Res18) [23]	74.8	74.7	65.5	GTX 1080Ti	1536×768	55.3	49M
BiSeNetV2-L [24]	75.8	75.3	47.3	GTX 1080Ti	1024×512	118.5	-
PP-LiteSeg-T2 [25]	76.0	74.9	**96.0**	RTX 3090	1536×768	-	-
PP-LiteSeg-B2 [25]	78.2	77.5	68.2	RTX 3090	1536×768	-	-
HyperSeg-M [26]	76.2	75.8	59.1	RTX 3090	1024×512	7.5	10.1M
HyperSeg-S [26]	78.2	78.1	45.7	RTX 3090	1536×768	17.0	10.2M
PIDNet-S [27]	78.8	78.6	93.2	RTX 3090	2048×1024	46.3	7.6M
PIDNet-M [27]	80.1	80.1	39.8	RTX 3090	2048×1024	197.4	34.4M
ESSNL-48 (Ours)	77.8	77.3	36.7	RTX 3090	2048×1024	183.5	4.8M
ESSNL-72 (Ours)	80.7	80.5	27.8	RTX 3090	2048×1024	263.7	7.0M
ESSNL-84 (Ours)	**81.3**	**81.2**	22.3	RTX 3090	2048×1024	304.5	8.1M

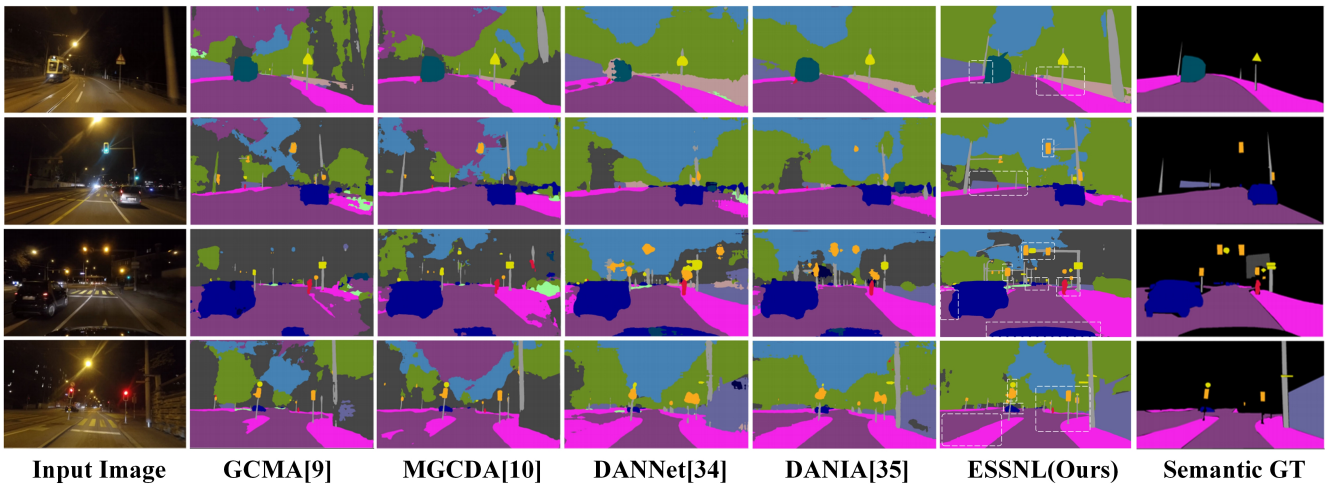


Fig. 5: Visualization comparison of our ESSNL with existing SOTA methods on four samples from Nighttime-Driving-test.

following datasets are used for model training and performance evaluation:

1) Cityscapes [28] has fine annotations for $2,975$ training images, 500 validation images, and $1,525$ test images. All images are at a fixed resolution of 2048×1024 pixels with pixel-level annotations of a total of 19 categories.

2) Dark Zurich [10] consists of $2,416$ nighttime images and $2,920$ twilight images, which are all unlabeled with a resolution of 1920×1080. It contains another 201 annotated nighttime images including 50 for validation (Dark-Zurich-

TABLE II: The Per-Category mIoU (%) on Dark Zurich-Test, where † indicates that only the cityscapes dataset is used for training.

Method	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	**mIoU**
RefineNet† [29]	68.8	23.3	46.8	20.8	12.6	29.8	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	28.5
PSPNet† [30]	78.2	19.0	51.2	15.5	10.6	30.3	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	28.8
AdaptSegNet† [31]	86.1	44.2	55.1	22.2	4.8	21.1	5.6	16.7	37.2	8.4	1.2	35.9	26.7	68.2	45.1	0.0	53.2	23.2	10.7	28.5
ADVENT† [32]	85.8	37.9	55.5	27.7	14.5	23.1	14.0	21.1	32.1	8.7	2.0	39.9	16.6	64.0	13.8	0.0	58.8	28.5	20.7	29.7
BDL† [33]	85.3	41.1	61.9	32.7	17.4	20.6	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	30.8
DMAda [8]	75.5	29.1	48.6	21.3	14.3	34.3	36.8	29.9	49.4	13.8	0.4	43.3	50.2	69.4	18.4	0.0	27.6	34.9	11.9	32.1
GCMA [9]	81.7	46.9	58.8	22.0	20.0	41.2	40.5	41.6	64.8	31.0	32.1	53.5	47.5	75.5	39.2	0.0	49.6	30.7	21.0	42.0
MGCDA [10]	80.3	49.3	66.2	7.8	11.0	41.4	38.9	39.0	64.1	18.0	55.8	52.1	53.5	74.7	**66.0**	0.0	37.5	29.1	22.7	42.5
DANNet (RefineNet) [34]	90.0	54.0	**74.8**	**41.0**	21.1	25.0	26.8	30.2	**72.0**	26.2	84.0	47.0	33.9	68.2	19.0	0.3	66.4	38.3	23.6	44.3
DANNet (PSPNet) [34]	90.4	60.1	71.0	33.6	22.9	30.6	34.3	33.7	70.5	31.8	80.2	45.7	41.6	67.4	16.8	0.0	73.0	31.6	22.9	45.2
DANIA (RefineNet) [35]	90.8	59.7	73.7	39.9	26.3	36.7	33.8	32.4	70.5	32.1	**85.1**	43.0	42.2	72.8	13.4	0.0	71.6	48.9	23.9	47.2
DANIA (PSPNet) [35]	91.5	**62.7**	73.9	39.9	25.7	36.5	35.7	36.2	71.4	**35.3**	82.2	48.0	44.9	73.7	11.3	0.1	64.3	36.7	22.7	47.0
ESSNL-84 (Ours)	**92.2**	61.6	72.0	35.9	**35.3**	**59.6**	**58.4**	**52.7**	60.0	23.8	68.0	**64.5**	**69.5**	**86.4**	11.8	**21.7**	**78.1**	**53.1**	**32.4**	**54.6**

val) and 151 for testing (Dark-Zurich-test).

3) Nighttime Driving [8] contains 50 nighttime images of resolution 1920×1080 from diverse visual scenes. All these 50 images have been annotated at the pixel level using the same 19 Cityscapes category labels.

Note that Dark-Zurich-test and Cityscapes-test serve as online benchmarks whose ground truths are not publicly available. For our experiments, we submit the segmentation results to the evaluation website of these datasets and subsequently obtain ESSNL's performance metrics.

3.2 Experimental Settings

Since ESSNL requires a labeled dataset and supervised training. 1) For Cityscapes testing, we utilize the provided trainning and validation set for training. 2) For Dark Zurich testing, we face challenges due to the limited number of annotated semantic segmentation datasets under low-light conditions. Given the inability to extract the model's optimal performance under low-light using only Cityscapes, we supplement the Cityscapes training and validation sets with Dark-Zurich-val and Nighttime-test at a 4:1 ratio. Following this, we conducted ESSNL training. 3) For Nighttime Driving testing, we use Cityscapes and supplemented Dark-Zurich-val for its training.

All the training in the experiments is carried out on a desktop with four NVIDIA RTX3090 graphics cards, each with 24G memory capacity on the Ubuntu system. We use the dual closed-loop training strategy introduced in Sec. II. We employ the SGD optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. For each model, we utilize the original image size of 2048×1024 as input, with an initial learning rate set to 10^{-4} and momentum fixed at 0.9. Additionally, we apply random cropping with a crop size of 512, within the scale range of 0.5 to 1.0, and random horizontal flipping. The batch size for training is set at 32, and the total iteration count is 50000.

3.3 Comparison With State-of-the-Art Methods

We first compare ESSNL with existing SOTA methods in real-time semantic segmentation via Cityscapes-Test, as reported in Table I. All of the three versions of ESSNL outperform the others. Since HGC selects a grouping dimension of 3, it utilize the least parameter increment to expand the receptive field, thus achieving a trade-off between accuracy and speed. Similar to these methods, ESSNL achieves real-time (30FPS) performance.

We next compare ESSNL with existing SOTA methods in semantic segmentation under low-light scenes via Dark-Zurich-test. As reported in Table II, ESSNL demonstrates the highest segmentation accuracy. Specifically, it surpasses MGCDA [10], DANNet [34], and DANIA [35], by 28.47%, 20.80%, and 15.68%, respectively, which is very significant. It achieves the highest accuracy across 12 out of 19 categories, including background categories like road surfaces, where there's a modest accuracy increase of 1.54% over DANIA [35]. Notably, in segmenting small targets like poles, pedestrians, and cars, it shows significant improvements of 63.30%, 34.38%, and 17.23%, respectively, over DANIA [35]. The expanded receptive field gradient of HGC can be the main reason for enhancing the segmentation accuracy of small object. Additionally, DCBM algorithm enables ESSNL to enhance the illuminated features of low-light images before decoding them, thus improving its segmentation accuracy under low-light scenes.

We finally compare ESSNL with SOTA methods via Nighttime-Driving-Test, as reported in Table V. In Fig. 4, we can observe that ESSNL significantly enhances road alignment accuracy with the semantic GT in rows 2 and 4 when compared to other methods. Rows 2, 3, and 4 show its accurate segmentation of traffic light and pole. In row 2, only ESSNL achieves the segmentation of building among these methods. It demonstrates minimal erroneously predicted pixels inside each classified area, which can be attributed to its masks-based decoder instead of pixel-based one.

3.4 Ablation Study

1) Adaptability to low-light images. To further reveal the capabilities of DCBM, we compare ESSNL that solely

TABLE III: Results of ESSNL with different layers tested on Cityscapes-val, where τ represents the number of processed frames per second on a single NVIDIA 3090 graphics card

Model	n	GFLOPs	Params	mIoU	iIoU	τ
ESSNL-48	6	183.5	4.8M	77.8	57.0	**36.7**
ESSNL-60	10	223.7	5.9M	79.5	58.4	32.0
ESSNL-72	14	264.5	7.0M	80.7	59.1	27.8
ESSNL-84	18	305.1	8.1M	81.3	**59.8**	22.3
ESSNL-96	22	345.3	9.2M	**81.4**	59.7	18.9

TABLE IV: Ablation Study on ESSNL on Dark-Zurich-val, where † represents methods used only Cityscapes for training

Method	mIoU
DANIA [35]	38.14
PIDNet-M† [27]	23.43
HyperSeg-S† [26]	24.65
ESSNL†(Ours)	36.25
ESSNL (Ours)	**45.37**
w/o Any HGC unit	41.81
w/o First 1/2 HGC units	42.25
w/o Last 1/2 HGC units	44.54
w/o Inner loop	28.60
w/o Dark Zurich -Train	44.03
w/o Nighttime Driving -Train	43.67
w/o Dark Zurich & Nighttime Driving -Train (on Dark-Zurich-test)	37.82
w/o Cityscapes -Train (on Dark-Zurich-test)	31.17

employs DCBM without HGC to other SOTA networks. All these networks are trained solely on Cityscapes, and then tested on Cityscapes-val and Dark-Zurich-val. Although these compared networks exhibit high segmentation performance on Cityscapes-val, their segmentation accuracy on Dark-Zurich-val significantly lags behind ESSNL, as reported in Table IV. In contrast, DCBM makes our ESSNL realize the synergy between a shallow illumination enhanced network and deep segmentation network, thus showing the segmentation adaptability under low-light conditions. The visualization in Fig. 4 illustrates ESSNL's superior adaptability to low-light environments over other networks. It demonstrates erroneously predicted pixels inside each classified area and the highest matching rate with the semantic GT.

2) Network Architecture. We performed experiments using various layers in ESSNL and evaluated their performance on the Cityscapes-val, as reported in Table III. Upon reaching 84 layers, we observed an increase in parameters, a drop in segmentation speed, and minimal accuracy enhancement. We thus selected ESSNL-84 as our optimal choice. Additionally, we investigated the effectiveness of HGC units by substituting them partially or completely with standard convolutions, yielding results outlined in Table IV. It demonstrates that HGC can improve the segmentation accuracy of ESSNL, especially when implemented in its shallow layers.

3) Training Strategy. To validate the efficacy of the proposed dual closed-loop loss optimization, we conduct training exclusively by using the outer loop, without involving the inner one. Additionally, we explore different training datasets, including using the Cityscapes only, incorporating labeled samples from only Dark-Zurich-val or Nighttime-Driving-test. The results in Fig. 5 indicate that training ESSNL with Cityscapes supplemented with other annotated dataset in low-light scenes can improve its segmentation accuracy. However, when training without Cityscapes, the performance of ESSNL on Dark-Zurich-val is significantly reduced, which may be due to the limited number of training sets. It showcases the advantage of ESSNL, which can be trained using existing datasets captured under day-light conditions and demonstrates adaptability for low-light scenes.

TABLE V: Hyperparameter sensitivity for Loss Weights in (6) on Dark-Zurich-val.

Weights	mIoU
β_1, β_2, β_3 = 1, 1, 1	15.70
β_1, β_2, β_3 = 0.1, 1, 1	42.23
β_1, β_2, β_3 = 1, 1, 1	17.09
β_1, β_2, β_3 = 0.1, 1, 0.1	**45.37**
β_1, β_2, β_3 = 0.01, 1, 0.1	43.85
β_1, β_2, β_3 = 0.1, 1, 0.01	44.02
β_1, β_2, β_3 = 0.01, 1, 0.01	25.27

TABLE VI: mIoU (%) results on Nighttime-Driving-Test.

Model	mIoU
GCMA [9]	45.60
MGCDA [10]	49.40
DANNet(RefineNet) [34]	42.36
DANNet(PSPNet) [34]	47.70
DANIA(RefineNet) [35]	45.65
DANIA(PSPNet) [35]	48.38
ESSNL-48(Ours)	51.20
ESSNL-72(Ours)	53.14
ESSNL-84(Ours)	**54.28**

4) Hyperparameter. To learn the sensitivity of our method to the choice of hyperparameters, we study the different choices of weights in (7) as shown in Table V. Due to the significant difference in loss magnitude between the shallow IE network's inner loop (L_1) and the network's outer loop Dice loss (L_2), we aimed to balance this by assigning smaller weights (β_1 and β_3) compared to β_2. We found that setting all weights to 0.1 achieved the highest testing accuracy.

IV. Conclusion

In this work, we introduce an End-to-end Semantic Segmentation Network for Low-light scenes (ESSNL). It addresses the challenge faced by existing low-light semantic segmentation methods that require the separate training of two models and thus result in performance disparities and a complex training process in practical applications [36]–[38]. Specifically, we propose HGC units to enhance the network's edge feature extraction capabilities, and DCBM to achieve end-to-end training for both illumination enhanced shallow network and deep semantic segmentation network. Tests on public datasets demonstrate that the segmentation accuracy of ESSNL well exceeds that of existing SOTA methods.

Despite being lightweight, HGC significantly increases time complexity over conventional convolution. Additionally, the limited availability of labeled semantic segmentation datasets for low-light scenes results in an imbalanced distribution of samples in our training set between dark and bright conditions. ESSNL thus may not have achieved its optimal performance yet. Recently developed domain adaptation methods, e.g., [39]–[46], should be investigated to address the limited labeled data issue.

REFERENCES

[1] Roggiolani *et al.*, "On domain-specific pre-training for effective semantic perception in agricultural robotics," in *2023 IEEE Int. Conf. on Robotics and Automation (ICRA)*, pp. 11786–11793, 2023.

[2] S. Kuutti *et al.*, "A survey of deep learning applications to autonomous vehicle control," *IEEE Trans. on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 712–733, 2021.

[3] C. Wang, W. Pedrycz, Z. Li, and M. Zhou, "Residual-driven fuzzy c-means clustering for image segmentation," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 4, pp. 876–889, 2021.

[4] C. Li *et al.*, "Low-light image and video enhancement using deep learning: A survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9396–9416, 2022.

[5] C. Li, C. Guo, and C. C. Loy, "Learning to enhance low-light image via zero-reference deep curve estimation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4225–4238, 2021.

[6] L. Ma *et al.*, "Toward fast, flexible, and robust low-light image enhancement," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 5637–5646, 2022.

[7] Y. Jiang *et al.*, "Enlightengan: Deep light enhancement without paired supervision," *IEEE Trans. on Image Processing*, vol. 30, pp. 2340–2349, 2021.

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

[9] C. Sakaridis, D. Dai, and L. V. Gool, "Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, pp. 7374–7383, 2019.

[10] C. Sakaridis, D. Dai, and L. Van Gool, "Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3139–3153, 2022.

[11] E. Romera *et al.*, "Bridging the day and night domain gap for semantic segmentation," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1312–1318, IEEE, 2019.

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

[13] S. Nag, S. Adak, and S. Das, "What's there in the dark," in *2019 IEEE Int. Conf. on Image Processing (ICIP)*, pp. 2996–3000, IEEE, 2019.

[14] H. Zhu and M. Zhou, "Efficient role transfer based on kuhnmunkres algorithm," *IEEE Trans. on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 42, no. 2, pp. 491–496, 2012.

[15] T. Cheng *et al.*, "Sparse instance activation for real-time instance segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 4433–4442, 2022.

[16] M. Weber, J. Luiten, and B. Leibe, "Single-shot panoptic segmentation," in *2020 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp. 8476–8483, IEEE, 2020.

[17] T. Takikawa *et al.*, "Gated-scnn: Gated shape cnns for semantic segmentation," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, pp. 5229–5238, 2019.

[18] Z. Cao, X. Xu, B. Hu, and M. Zhou, "Rapid detection of blind roads and crosswalks by using a lightweight semantic segmentation network," *IEEE Trans. on Intelligent Transportation Systems*, vol. 22, no. 10, pp. 6188–6197, 2020.

[19] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact++: Better real-time instance segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2020.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf, on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[21] M. Orsic, I. Kreso, P. Bevandic, and S. Segvic, "In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 12607–12616, 2019.

[22] S. Kumaar, Y. Lyu, F. Nex, and M. Y. Yang, "Cabinet: Efficient context aggregation network for low-latency semantic segmentation," in *2021 IEEE Int. Conf. on Robotics and Automation (ICRA)*, pp. 13517–13524, IEEE, 2021.

[23] C. Yu *et al.*, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. of the European Conf. on Computer Vision (ECCV)*, pp. 325–341, 2018.

[24] C. Yu *et al.*, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. Journal of Computer Vision*, vol. 129, pp. 3051–3068, 2021.

[25] J. Peng *et al.*, "Pp-liteseg: A superior real-time semantic segmentation model," *arXiv preprint arXiv:2204.02681*, 2022.

[26] Y. Nirkin *et al.*, "Hyperseg: Patch-wise hypernetwork for real-time semantic segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 4061–4070, 2021.

[27] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "Pidnet: A real-time semantic segmentation network inspired by pid controllers," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 19529–19539, 2023.

[28] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3213–3223, 2016.

[29] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1925–1934, 2017.

[30] H. Zhao *et al.*, "Pyramid scene parsing network," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2881–2890, 2017.

[31] Y.-C. Chen *et al.*, "Crdoco: Pixel-level domain transfer with cross-domain consistency," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 1791–1800, 2019.

[32] T.-H. Vu *et al.*, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 2517–2526, 2019.

[33] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 6936–6945, 2019.

[34] X. Wu *et al.*, "Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 15769–15778, 2021.

[35] X. Wu, Z. Wu, L. Ju, and S. Wang, "A one-stage domain adaptation network with image alignment for unsupervised nighttime semantic segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 58–72, 2023.

[36] H. Mu *et al.*, "Dynamic obstacle avoidance system based on rapid instance segmentation network," *IEEE Trans. on Intelligent Transportation Systems*, pp. 1–15, 2023.

[37] X. Wang *et al.*, "Domain adaptation multitask optimization," *IEEE Trans. on Cybernetics*, vol. 53, no. 7, pp. 4567–4578, 2023.

[38] X. Guo, W. Zhou, and T. Liu, "Contrastive learning-based knowledge distillation for rgb-thermal urban scene semantic segmentation," *Knowledge-Based Systems*, p. 111588, 2024.

[39] G. Cai *et al.*, "Unsupervised domain adaptation with adversarial residual transform networks," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 31, no. 8, pp. 3073–3086, 2020.

[40] S. Teng *et al.*, "Adaptive graph embedding with consistency and specificity for domain adaptation," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 11, pp. 2094–2107, 2023.

[41] F. Aimedee *et al.*, "Systematization of morphing in reconfigurable mechanisms," *Mechanism and machine theory*, vol. 96, pp. 215–224, 2016.

[42] Q. Kang *et al.*, "Effective visual domain adaptation via generative adversarial distribution matching," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 32, no. 9, pp. 3919–3929, 2021.

[43] G. Ma *et al.*, "Estimating the state of health for lithium-ion batteries: A particle swarm optimization-assisted deep domain adaptation approach," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 7, pp. 1530–1543, 2023.

[44] Q. Kang *et al.*, "Enhanced subspace distribution matching for fast visual domain adaptation," *IEEE Trans. on Computational Social Systems*, vol. 7, no. 4, pp. 1047–1057, 2020.

[45] Z. Zheng *et al.*, "Knowledge transfer learning via dual density sampling for resource-limited domain adaptation," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 12, pp. 2269–2291, 2023.

[46] S. Yao *et al.*, "Discriminative manifold distribution alignment for domain adaptation," *IEEE Trans. on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 2, pp. 1183–1197, 2023.