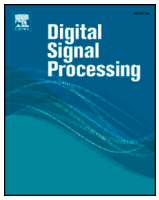
Digit. Signal Process. 155 (2024) 104752

Contents lists available at ScienceDirect

Digital Signal Processing

journal homepage: www.elsevier.com/locate/dsp



# Lane detection based on real-time semantic segmentation for end-to-end autonomous driving under low-light conditions

Yang Liu [a], Yongfu Wang [a,*], Qiansheng Li [b]

[a] School of Mechanical Engineering and Automation, Northeastern University, Shenyang 110819, China
[b] Dalian Power Plant, HUANENG Power Intl Inc, Dalian 116100, China

ARTICLE INFO

*Keywords:*
End-to-end
Real-time
Semantic segmentation
Autonomous driving system
STDC
Image enhancement

ABSTRACT

Lane detection is one of the important modules of the autonomous driving system for environmental perception. In recent years, lane detection based on the semantic segmentation method has effectively promoted the development of autonomous driving technology. It provides lane information of complex driving scenes for the Deep Reinforcement Learning (DRL) method as a decision-making reference, which is the key technology to achieve vehicle-assisted driving and even autonomous driving. Aiming at the problems of feature extraction difficulty and low utilization rate of semantic feature information in lane detection method based on semantic segmentation in low-light scenes, which lead to miss-detection or false-alarm, this study designs a real-time lane segmentation method based on the encoding-decoding mechanism in low-light scenes, namely UET-STDC. The method first uses the ZERO-DCE++ method to improve the quality of low-light road video frames and then designs a multi-scale feature extraction module combined with deformable convolution to improve the method's ability to extract lane semantic features in complex environments and uses skip connections to promote the fusion of shallow semantic features and deep semantic features. Finally, some of the up-sampling modules in the encoder structure are simplified, and the depthwise separable convolution is introduced to replace the convolution to reduce the computational complexity of the UET-STDC method. In addition, in order to obtain a better representation of lane features, this study combines attention mechanisms in the encoder structure. A large number of experiments have been carried out on the self-made lane semantic segmentation dataset. The results show that the UET-STDC method can significantly improve the accuracy of lane segmentation under low-light conditions and has a good segmentation effect and robustness under various illumination conditions.

## 1. Introduction

As an essential part of the Advanced Driving Assistance System (ADAS) and the autonomous driving system, lane detection is of great significance in ensuring the safety of autonomous driving [1]. However, due to the low-illumination of images or video frames acquired by on-board optical sensors under low-light conditions, the complexity of road scene illumination, and the impact of tree shadows, road cracks, and other factors, the accuracy of the lane detection method is low [2]. In the autonomous driving system based on the DRL method, the quality of the lane semantic segmentation method significantly affects the decision-making ability. Low-quality semantic segmentation results will lead to the problem of autonomous driving vehicles not being able to brake in time and change lanes normally. At the same time, if the vehicle cannot reach the millisecond-level response at high speed, the overall safety of the autonomous driving system will be affected [3,4]. Lanes have long and thin appearance structures and are vulnerable to vehicle occlusion and shadow occlusion. Moreover, the shape of the lane is not fixed, and the degree of road curvature will also affect the semantic segmentation performance [5]. When the illumination conditions are poor, the impact of the above problems on the semantic segmentation method becomes more serious, and the lane will become very fuzzy, which will make it difficult for the semantic segmentation method to extract the semantic feature information of the lane effectively, and thus can not accurately identify the lane. In addition, the labels in mainstream lane datasets are usually sparse, which affects the semantic segmentation methods' ability to sense the road scene's structure [6]. Real-time semantic segmentation methods are widely used in Computer Vision (CV) tasks such as autonomous driving [7]. However, due to the limitations of storage space and hardware performance, it is still a considerable

* Corresponding author.
*E-mail address:* yfwang@mail.neu.edu.cn (Y. Wang).

https://doi.org/10.1016/j.dsp.2024.104752

Available online 5 September 2024
1051-2004/© 2024 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

challenge to store and run real-time semantic segmentation models on vehicle-embedded terminals. In order to solve the above problems, Fan et al. [8] proposed the STDC method based on BiSeNetV1 [9]. It deleted the dual-branch network structure in the BiSeNetv1 network and used a new training method to improve segmentation performance without increasing any reasoning cost.

The low-light image enhancement method is currently a popular image processing technology, which can be classified into traditional methods and the Deep Learning (DL) method [10]. However, most of the traditional low-light image enhancement methods only focus on improving the contrast or the illumination of the image, ignoring the impact of noise, which seriously affects the enhancement effect. In recent years, researchers have proposed many methods based on the DL method for enhancing low-light images. The method for enhancing low-light images based on DL has become a prominent means of low-light image enhancement. Currently, many research works have used Convolutional Neural Network (CNN) as the basis of low-light image enhancement methods based on DL. For example, Tao et al. [11] proposed a low-light convolutional neural network method for low-light image enhancement. The network uses the CNN method to extract multi-scale features of the image and uses structural similarity as the loss function to constrain the mapping relationship between the real image and the low illumination image to enhance the image. Li et al. [12] proposed a trainable CNN for low-light image enhancement, namely LightenNet, which takes the low-light image as input and outputs its illumination map, then uses the illumination map to input the network model based on Retinex to obtain the enhanced image. Chen et al. [13] established a dataset named See-in-the-Dark (SID) containing original low-exposure and low-illumination images, as well as corresponding long-exposure images, and proposed a low-light image enhancement Neural Network (NN) structure based on the Full Convolution Network (FCN). Wang et al. [14] proposed a novel Deep Lightening Network (DLN) that consists of several Lightening Back-Projection (LBP) blocks and a Feature Aggregation (FA) block, and the experiment shows that the DLN method outperforms other methods under both objective and subjective metrics. Images captured in low-light environments often suffer from issues related to low illumination and damaged details, which results in poor visibility. Hu et al. [15] proposed a new end-to-end neural network architecture for low-light remote-sensing low-light image enhancement, named Remote-Sensing CNN (RSCNN). The RSCNN method improves the visibility of the image. The low-light level image enhancement method based on DL often has a complex structure and brings a huge computational burden, which hinders its deployment on vehicle-borne embedded terminals. To solve this problem, Li et al. [16] proposed a lightweight and efficient Luminance-aware Pyramid Network (LPNet) to reconstruct normal-illumination images in a coarse-to-fine strategy. Li et al. [17] proposed a practical real-time low-light image enhancement network named Zero-DCE++. Experimental results demonstrate the effectiveness and practicality of the Zero-DCE++ method on various datasets. Therefore, we introduce the Zero-DCE++ method in this paper to enhance the autonomous driving environment image under various illumination environments by enhancing the dark area while maintaining the bright area of the image.

Lane detection under low-light conditions mainly faces the following three challenges: 1) Due to poor perception ability under low-light conditions, it is difficult to obtain large-scale annotated road scene datasets under low-light conditions. Deep learning is essentially a data-driven approach. At present, the standard strategy for achieving high-performance semantic segmentation is to use neural networks to train a large amount of annotated low-light image data. The process of collecting and annotating low-light images is labor-intensive. Due to the large area of darkness or shadows in images under low-light conditions, it is very difficult to establish high-quality annotations at the pixel level. 2) There are problems with low brightness, noise, and motion blur in low-light conditions road scene images, and there are significant differences between the features extracted using convolutional layers and those obtained under good lighting conditions. Therefore, models trained on road datasets under normal lighting cannot be directly applied to road scenes under low-light conditions. 3) Compared with the BiSeNetV1 method, STDC is very suitable for onboard computer environments. However, the STDC method still has some problems, such as low segmentation accuracy and poor network convergence. At the same time, in the low-light environment, the image definition is low, the edge is blurred, and the image details and texture structure are not well preserved, which seriously affects the segmentation effect of STDC. Therefore, we combine the real-time low-light image enhancement method based on DL with the road scene semantic segmentation method based on UET-STDC to improve the real-time and reliability of automatic driving and reduce the hardware and software costs of automatic driving. The following is a summary of this paper's contributions:

- The Zero-DCE++ method for low-light image enhancement is introduced to improve the quality of images and get over the model's performance bottleneck in low-light image enhancement while still providing high-quality images for image processing and analysis. In low-light environments, this method obtains more precise semantic information while maintaining the visual look compared to the previously proposed DL methods.
- An autonomous driving system named UET-STDC based on lane semantic segmentation and DRL is designed. The UET-STDC method is proposed to segment the lane of the image captured by the vehicle camera, eliminate the interference of environmental noise, and only retain the key information of the lane in the original image so as to provide marked video frames for the autonomous driving decision-making module based on DRL.
- Some problems exist in the current public lane semantic segmentation dataset, such as image data redundancy, outdated lane types, and the unbalanced distribution of lane images. In order to solve these problems, this study first collected road images under various illumination conditions in multiple cities in China through vehicle-mounted cameras, then systematically constructed the lane semantic segmentation dataset and expanded the dataset through rotation, random erasure, and other methods. The lane semantic segmentation dataset has more significant advantages in image number, lane type, and scalability than existing datasets.
- The semantic segmentation method based on the pure vision in the low-light environment proposed in this study has better practicability and economy than that based on laser radar or infrared thermal imager in the low-light environment and can meet the real-time and accuracy requirements of semantic segmentation technology for vehicle mobile terminals and vehicle embedded devices. At the same time, in order to better provide inspiration for relevant researchers, we discussed the possibility of combining the UET-STDC semantic segmentation method with the DRL.

Following are the key points of each section of this essay: Section 1 is an introduction that mainly describes the background of this study, traditional low-light image enhancement methods, low-light image enhancement methods, and the study status of end-to-end autonomous driving technology, and introduces the study purpose, main work and section content arrangement of this paper. Section 2 is the theoretical basis and the network structure of the UET-STDC method proposed in this paper. This chapter introduces the relevant theoretical basis of this paper in detail, including the ZERO-DCE++ image enhancement method, attention mechanism, and the establishment process of the self-made lane semantic segmentation dataset. Section 3 introduces the preparation of the experiment and how to demonstrate the effectiveness of the UET-STDC method through a series of comparative experiments. Section 4 introduces the ablation experiment and discussion designed in this paper. This chapter discusses in detail how we demonstrate the impact of each innovation point of this study on performance through ablation experiments and how the UET-STDC method is combined with the DRL

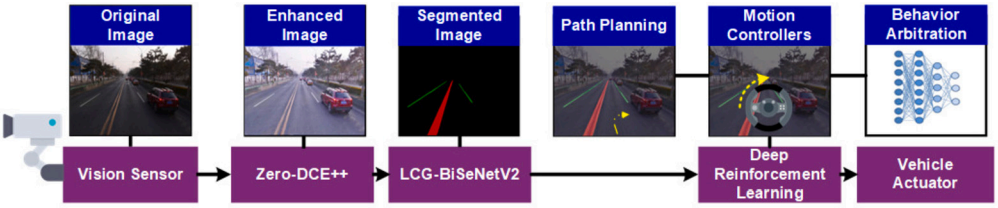



**Fig. 1.** Image acquisition, processing, and autonomous driving decision-making process: autonomous driving vehicles collect the traffic state image in front of the vehicle through the vehicle-borne optical imaging sensor system, the UET-STDC method returns the processed image, and the DRL method outputs autonomous driving decision-making through the processed image.

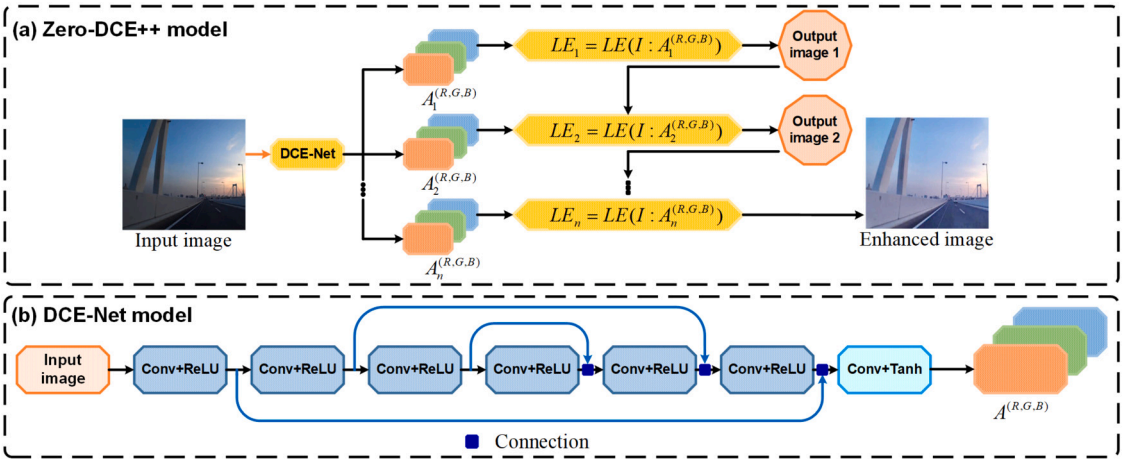



**Fig. 2.** (a) The framework of the Zero-DCE++ method. (b) The framework of the DCE-Net module. The DCE-Net module is devised to estimate a set of best-fitting Light-Enhancement curves (LE-curves) that iteratively enhance a given input image (i.e., takes the enhanced image as the input of the next iteration, and the input is enhanced in a progressive manner).

to apply in automatic driving. Section 5 is the summary and outlook of this paper. This section summarizes the research results of this paper and looks forward to future work.

## 2. The methods

In this study, the semantic segmentation of the automatic driving environment is based on the UET-STDC method trained with the self-made lane semantic segmentation dataset. The labeled video frames output by the UET-STDC method are regarded as the input state of the DRL method. This section mainly introduces the background and component modules of the UET-STDC method.

### 2.1. The Zero-DCE++ model

The Zero-DCE++ method is an important part of the LLIE-LiteSeg-T1 method. The LLIE-LiteSeg-T1 method can be combined with the reinforcement learning method to be applied in the real autonomous driving environment (see Fig. 1). Inspired by the Retinex model, the Zero-DCE++ method designed a new low-light image enhancement method, which does not directly perform image-to-image enhancement mapping but redefines the low-light image enhancement task as the illumination mapping curve estimation problem. The image is enhanced by the estimated illumination mapping curve to speed up the algorithm's execution, and the speed of image processing can reach 1000 FPS. The algorithm structure of the Zero-DCE++ method is shown in Fig. 2.

In order to achieve the goal of image enhancement, the Zero-DCE++ method has designed a conic curve named LE-curve. The conic curve is modeled as follows:

$$LE(I(\mathbf{x});\alpha) = I(\mathbf{x}) + \alpha I(\mathbf{x})(1 - I(\mathbf{x})), \tag{1}$$

where $x$ represents pixel coordinates, $LE((I((x);\alpha)$ is an enhanced version of the given input $I(x)$, which can train curve parameters $\alpha \in [-1, 1]$ and adjust the shape of the LE-curve and control the exposure level.

The LE-curve defined in function (1) can be applied iteratively to enable more useful adjustments to deal with challenging low-light environments. The iteratively LE-curve formula can be expressed as follows:

$$LE_n(\mathbf{x}) = LE_{n-1}(\mathbf{x}) + \alpha_n LE_{n-1}(\mathbf{x})\left(1 - LE_{n-1}(\mathbf{x})\right), \tag{2}$$

where $n$ is the number of iterations that control the curvature; according to the research [17], we set $n = 8$, which can handle most cases well.

In order to achieve local adjustment, the single parameter of the higher-order curve $\alpha$ can be changed to pixel-level parameters to get pixel-level curves. Each pixel of a given input image has the best fit $\alpha$. The corresponding curve is used to adjust its dynamic range. Therefore, the curve formula is redefined as:

$$LE_n(\mathbf{x}) = LE_{n-1}(\mathbf{x}) + \mathcal{A}(\mathbf{x})LE_{n-1}(\mathbf{x})\left(1 - LE_{n-1}(\mathbf{x})\right), \tag{3}$$

where $\mathcal{A}(\mathbf{x})$ is a parameter mapping of the same size as the given image, we only estimate three curve parameter mappings and then reuse them in different iteration stages.

The DCE-Net network structure diagram is shown in Fig. 2. The input is a low-light image, and the output is a pixel-level curve parameter mapping of a group of corresponding high-order curves. In this paper, skip connection with seven Depthwise Separable Convolution (DSC) layers is adopted, and each layer contains 32 CNNs, the size of convolution kernel is $3 \times 3$ and the stride of convolution kernel is 1, and then ReLU activation function is used. Behind the last DSC layer is the Tanh activation function, which generates 24 feature maps ($n = 8$) for 8 iterations, where each iteration generates 3 curve feature maps for RGB channels.

#### 2.1.1. The experiment and analysis of image quality assessment

1. The Natural Image Quality Evaluator (NIQE) method [22]. Because it is impossible to obtain completely original undistorted images as

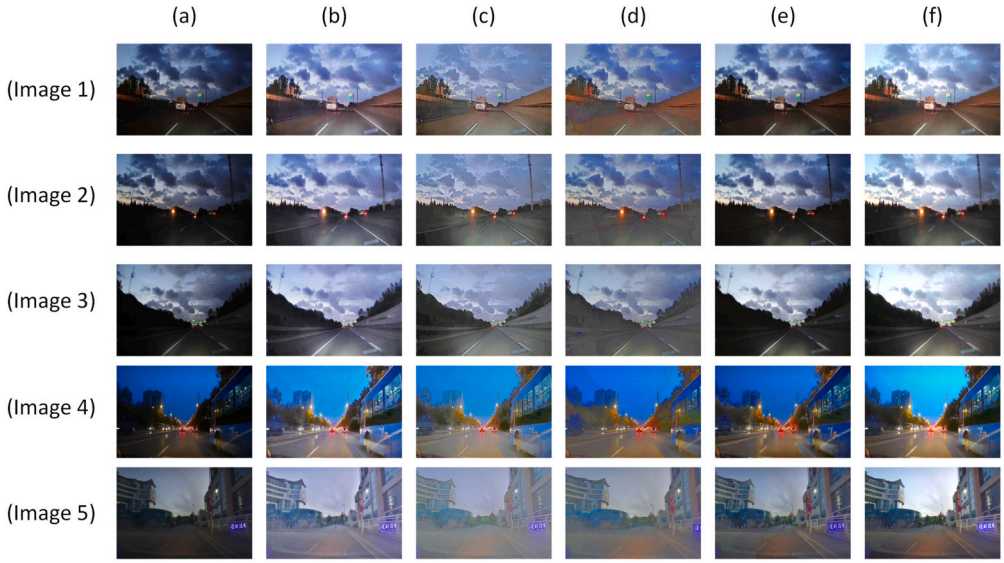



**Fig. 3.** Comparison of the enhancement effect between Zero-DCE++ and several mainstream image enhancement methods. (a) The initial images. (b) Zero-DCE++ [17]. (c) RRDNet [18]. (d) MBLLEN [19]. (e) LLNet [20]. (f) DLN [21].

the reference image in the real environment, and the NIQE method does not rely on the information of the original undistorted image, the NIQE method has broad application prospects. It can be used as a method to assess the enhanced image quality of autonomous driving scenes. The normalized illumination $\hat{I}(i,j)$ of the image is calculated by the normalization method. We assumed that the illumination of the image is $I(i,j)$; the normalization calculation formula is as follows:

$$\sigma(i,j)=\sqrt{\sum_{y=-Y}^{Y}\sum_{t=-T}^{L}w_{y,t}\left(I_{y,t}(i,j)-\mu(i,j)\right)^2}, \tag{4}$$

$$\mu(i,j)=\sum_{k=-Y}^{Y}\sum_{l=-T}^{T}\omega_{y,t}I_{y,t}(i,j) \tag{5}$$

$$\hat{I}(i,j)=\frac{I(i,j)-\mu(i,j)}{\sigma(i,j)+1}, \tag{6}$$

where $i$ and $j$ are spatial domain coefficients, and $i\in 1,2,\cdots,\varpi$, $j\in 1,2,\cdots,\Lambda$, where $\varpi$ and $\Lambda$ are the height and width of the image respectively; $w=\left\{w_{y,t}\,|y=-Y,\cdots Y,t=-T,\cdots,T\right\}$ is gaussian kernel.

In this article, we divide the image into image block B of size $P\times P$ so as to calculate the local average variance of image block B:

$$\sigma(b)=\sum\sum_{(i,j)\in B}\sigma(i,j), \tag{7}$$

where $\sigma(b)$ is the local average variance of the image block B.

We use the Generalized Gaussian Distribution (GGD) [23] method to fit the normalized coefficients of image blocks. The product of four adjacent coefficients is fitted with the Asymmetric Generalized Gaussian Distribution (AGGD) [24], and then 16 parameters of adjacent coefficients are obtained by using the fast matching method, and the extracted features are used to fit the MVG model to calculate the parameters $\upsilon$ and $\Xi$. The formula of the MVG model is as follows:

$$f_x\left(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_k\right)=\frac{1}{(2\pi)^{k/2}|\Xi|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\upsilon)^{\mathrm{T}}\Xi^{-1}(\boldsymbol{x}-\upsilon)\right), \tag{8}$$

where $\left(x_1,x_2,\cdots,x_k\right)$ is the extracted image feature, $\upsilon$ and $\Xi$ can be obtained by maximum likelihood estimation.

Finally, we extracted the features of the distorted image and fit the MVG model to obtain $(\upsilon^{\varsigma},\Xi^{\varsigma})$. The image quality is measured by calculating the distance between the distorted image and the fitting parameters of the natural image. The specific calculation formula is as follows:

$$D\left(\boldsymbol{v}_1,\boldsymbol{v}_2,\Xi_1,\Xi_2\right)=\sqrt{\left(\boldsymbol{v}_1-\boldsymbol{v}_2\right)^{\mathrm{T}}\left(\frac{\Xi_1+\Xi_2}{2}\right)^{-1}\left(\upsilon_1-\upsilon_2\right)}, \tag{9}$$

where $\left(\upsilon_1,\upsilon_2,\Xi_1,\Xi_2\right)$ are the MVG model mean matrix and variance matrices of natural images and distorted images, respectively. The larger the values of these parameters, the better the image quality.

2. The Information Entropy (IE) method.
This research employs IE to objectively evaluate the improved image in order to understand the small variations in the image more logically. Entropy is an estimate of the amount of information that will be available before the result is known. Information is the information provided by a given event, and entropy is an estimate of the amount of information that will be available before the result is known. The average quantity of information in an image is reflected by image information entropy, which is a statistical form of feature [25]. The image's IE is then represented as follows:

$$H=-\sum_{i=1}^{L}p\left(a_i\right)\log_2 p\left(a_i\right), \tag{10}$$

where $a_i$ is the random output signal of the image.

3. The Contrast-changed Image Quality (CEIQ) method [26].
The CEIQ method has proposed a very simple but effective measurement index to predict the quality of contrast-distorted images. Its design is based on the fact that a high-contrast image is usually more similar to the contrast-enhanced image of the image. Specifically, firstly, HE is used to generate an enhanced image of the image to be tested, and then the structural similarity index is used as the first feature to calculate the similarity between the image to be tested and its enhanced image.

We selected and tested five low-light images to verify the performance of the proposed model, as shown in Fig. 3. In visual comparison, it can be seen that the Zero-DCE++ method can significantly enhance low-visibility images. As shown in Table 1, it can be seen that compared with

**Table 1**
Comparison with mainstream approaches.

| Method | Indicator | Image1 | Image2 | Image3 | Image4 | Image5 | Average value |
|---|---|---|---|---|---|---|---|
| Zero-DCE++ | NIQE↓ | 5.9595 | 5.6834 | 6.0062 | 5.3397 | 5.5921 | 5.7162 |
| | IE↑ | 7.7566 | 7.7672 | 7.7813 | 7.5989 | 6.7285 | 7.5325 |
| | CEIQ↑ | 3.5917 | 3.5777 | 3.6013 | 3.5367 | 3.2093 | 3.5033 |
| RRDNet | NIQE↓ | 6.3052 | 6.1701 | 6.6100 | 5.7158 | 5.4109 | 6.0424 |
| | IE↑ | 7.5118 | 7.5092 | 7.5537 | 7.4368 | 6.5528 | 7.3129 |
| | CEIQ↑ | 3.3872 | 3.3466 | 3.4522 | 3.3641 | 2.7834 | 3.2667 |
| MBLLEN | NIQE↓ | 6.2344 | 5.7332 | 6.3202 | 5.8133 | 6.6053 | 6.1413 |
| | IE↑ | 7.4654 | 7.0209 | 7.1583 | 7.2218 | 6.6895 | 7.1112 |
| | CEIQ↑ | 3.2003 | 3.1418 | 3.2953 | 3.2083 | 2.9330 | 3.1557 |
| LLNet | NIQE↓ | 6.4318 | 5.5816 | 6.0429 | 5.4539 | 4.4708 | 5.5962 |
| | IE↑ | 7.4659 | 7.4499 | 7.5306 | 7.3944 | 6.6526 | 7.3587 |
| | CEIQ↑ | 3.4267 | 3.4293 | 3.4802 | 3.3999 | 3.1817 | 3.3837 |
| DLN | NIQE↓ | 6.0562 | 5.7261 | 6.0442 | 5.1595 | 3.5621 | 5.3096 |
| | IE↑ | 7.6924 | 7.6706 | 7.3685 | 7.2512 | 6.3587 | 7.2683 |
| | CEIQ↑ | 3.5969 | 3.6699 | 3.5706 | 3.5166 | 3.2979 | 3.5304 |

**Table 2**
For the detailed structure of STDC, note that ConvX, as shown in the table, refers to the Conv-BN-ReLU. KSize means kernel size. S, R, and C denote stride, repeat times, and output channels, respectively.

| Stage | Output size | KSize | S | STDC1 | | STDC2 | |
|---|---|---|---|---|---|---|---|
| | | | | R | C | R | C |
| Input | 224 × 224 | | | | 3 | | 3 |
| ConvX1 | 112 × 112 | 3 × 3 | 2 | 1 | 32 | 1 | 32 |
| ConvX2 | 56 × 56 | 3 × 3 | 2 | 1 | 64 | 1 | 64 |
| Stage3 | 28 × 28 | | 2 | 1 | 256 | 1 | 256 |
| | 28 × 28 | | 1 | 1 | 256 | 3 | 256 |
| Stage4 | 14 × 14 | | 2 | 1 | 512 | 1 | 512 |
| | 14 × 14 | | 1 | 1 | 512 | 4 | 512 |
| Stage5 | 7 × 7 | | 2 | 1 | 1024 | 1 | 1024 |
| | 7 × 7 | | 1 | 1 | 1024 | 2 | 1024 |
| ConvX6 | 7 × 7 | 1 × 1 | 1 | 1 | 1024 | 1 | 1024 |
| GlobalPool | 1 × 1 | 7 × 7 | | | | | |
| FC1 | | | | | 1024 | | 1024 |
| FC2 | | | | | 1000 | | 1000 |

**Table 3**
Comparison between the mainstream semantic segmentation methods on the Cityscapes dataset.

| Model | Backbone | Resolution | FPS | mIoU(%) |
|---|---|---|---|---|
| ENet | no | 512 × 1024 | 76.9 | 58.4 |
| DABNet | no | 1024 × 2048 | 27.7 | 70.1 |
| CAS | no | 768 × 1536 | 108.2 | 70.6 |
| ICNet | PSPNet50 | 1024 × 2048 | 30.3 | 69.5 |
| GAS | no | 768 × 1536 | 108.4 | 71.2 |
| SFNet | DF1 | 1024 × 2048 | 119.6 | 73.3 |
| HMSeg | no | 768 × 1536 | 83.0 | 73.0 |
| BiSeNetV2 | no | 512 × 1024 | 154.2 | 72.9 |
| STDC1-Seg50 | STDC1 | 512 × 1024 | 250.4 | 71.9 |
| STDC2-Seg50 | STDC2 | 512 × 1024 | 188.6 | 73.4 |
| STDC1-Seg75 | STDC1 | 768 × 1536 | 126.7 | 75.3 |
| STDC2-Seg75 | STDC2 | 768 × 1536 | 97.0 | 76.8 |

several mainstream image enhancement algorithms, the Zero-DCE++ method can effectively improve the contrast of low-light images.

### 2.2. STDC method

In the current real-time semantic segmentation methods, DFANet [27] and BiSeNetv1 [9] make up for the decline in accuracy by using real-time backbone networks. However, these real-time backbone networks are designed for image classification and cannot meet the requirements of semantic segmentation well. One way to ensure real-time is to reduce the size of the input image, but the small size image ignores the object's boundary, details, and texture. In order to solve these problems, BiSeNetv2 [28] uses a dual-branch framework to combine detailed information and semantic information. However, the dual-branch framework reduces the reasoning speed of semantic segmentation methods and makes it challenging to extract low-level features effectively. These reasons significantly affect the segmentation accuracy of semantic segmentation methods. Therefore, the STDC method (see Fig. 4) is proposed. First, a new structure called the Short-Term Dense Concatenate (STDC) module is designed to obtain receptive fields of different sizes. Then, the STDC module is integrated into the U-net architecture to build the STDC network, thus significantly improving the semantic segmentation performance. The STDC method removes the spatial path from BiSeNetv2 and uses the detail guidance structure to replace the function of the spatial path. The detail guidance structure uses the Laplace kernel to generate a detailed ground-truth generation and then uses the detail ground-truth generation and the eight times down-sampled feature graph to calculate the detail loss. Improve the ability of network learning edge information and texture information through backpropagation of detail loss. STDC uses the detail guidance module to extract more spatial features. First, the detail aggregation module is used to generate detailed ground truth. Then, binary cross entropy loss and dice loss are used to optimize the detailed information learning process. Finally, semantic segmentation results are predicted by integrating the underlying spatial details and semantic information. The attention mechanism is introduced into the Feature Fusion Module (FFM) and Attention Refinement Module (ARM) to obtain high-precision semantic segmentation results. The detailed structure of BiSeNetV2 is shown in the Table 2. After these improvements, the precision of STDC on the cityscapes dataset has been significantly improved (see Table 3).

### 2.3. SwinT module

The road traffic environment of autonomous driving is very complex and changeable, and there are a lot of redundant objects or lack of visibility. Due to the powerful performance ability of the transformer

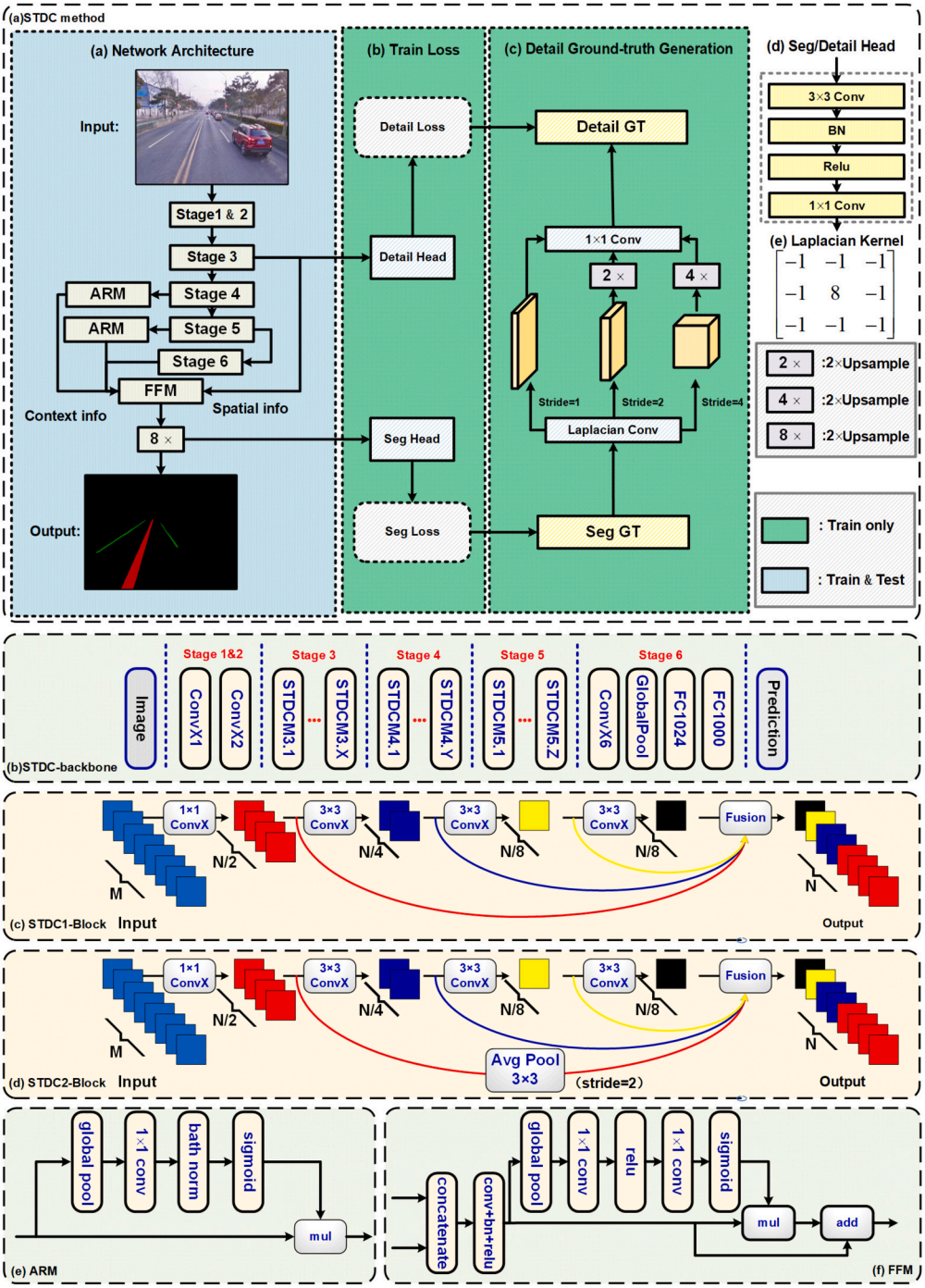



**Fig. 4.** The network structure of the STDC method, the STDC backbone, and the STDC-1 block. The first to fifth stages of the STDC method are the same as those of the STDC backbone.

method, researchers have proposed many methods to apply the transformer to semantic segmentation tasks. They have achieved good results, even greatly exceeding the CNN method. However, the visual entities of autonomous driving scenes change greatly, and the performance of the transformer method in different scenes is difficult to meet the actual needs. The autonomous driving scene image has high resolution and so many pixels. The transformer's computation based on non-local self-attention leads to much computation. The Swin-Transformer (SwinT) module can deal with these problems well [29]. The backbone network of SwinT is shown in Fig. 5. The SwinT module proposed the Windows Multi-head Self-Attention (W-MSA) (see Fig. 5). The Shifted Windows Multi-head Self-Attention (SW-MSA) (see Fig. 5) greatly reduces the amount of computation. Therefore, the global and local self-attention mechanisms can be introduced simultaneously to improve the feature extraction ability. SwinT's strategy of controlling the computing area in the patch greatly reduces the computational load of the network and the complexity of the linear proportion of the image size.

In Fig. 5, the input feature $\hat{z}^l$ of the Swin Transformer Block based on W-MSA is normalized through the Layer Normalization (LN) layer, then feature learning is conducted through the W-MSA module, and then the residual operation is performed to obtain the $z^l$. Then, the output feature $\hat{z}^{l+1}$ is obtained through an LN layer, a Multi-Layer Perceptron (MLP) layer, and a residual module. The input feature of the SwinT Block based

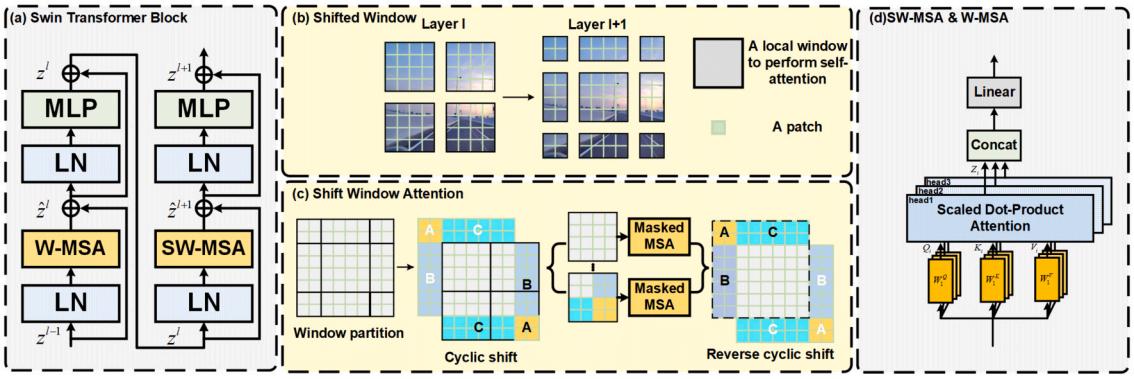



**Fig. 5.** The main components and steps of the SwinT method.

on SW-MSA is $z^{l+1}$. The calculation method of two consecutive SwinT Blocks is as follows:

$$\hat{z}^l = \text{W-MSA}\left(\text{LN}\left(z^{l-1}\right)\right) + z^{l-1}. \tag{11}$$

$$z^l = \text{MLP}\left(\text{LN}\left(\hat{z}^l\right)\right) + \hat{z}^l. \tag{12}$$

$$\hat{z}^{l+1} = \text{SW-MSA}\left(\text{LN}\left(z^l\right)\right) + z^l. \tag{13}$$

$$z^{l+1} = \text{MLP}\left(\text{LN}\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1}. \tag{14}$$

The self-attention mechanism is the key module of the SwinT method (see Fig. 5). The following stages are involved in calculating the attention value: To derive the weight, first compute the similarity between the generated attention-related query vector $Q$ and each key vector $K$. Secondly, the obtained weights are normalized by using the softmax function. Finally, the weight and the corresponding value vector $V$ are weighted and summed to obtain the final attention value. The MSA module splices several calculated self-attention heads in series, called Multi-heads, that can be expressed as:

$$\text{MultiHead}\,(Q,K,V) = \text{Concat}\left(\text{head}_1, \text{head}_2, \ldots, \text{head}_h\right) W^{\text{out}}, \tag{15}$$

where the Concat refers to the splice operation, and $W^{out}$ refers to the matrix for linear transformation. The initial self-attention result achieved by utilizing the scale's dot product attention is $headh$, which is written as follows:

$$\text{head}_h = \text{Attention}\left(Q_h, K_h, V_h\right) = \text{softmax}\left(\frac{Q_h \cdot K_h^T}{\sqrt{d_k}}\right) V_h, \tag{16}$$

The variance of the dot product of $Q_h$ and $V_h$ alleviates the gradient disappearance problem of softmax. Vectors $Q_h$, $K_h$ and $V_h$ are given as follows:

$$Q_h = IW_h^Q. \tag{17}$$

$$K_h = IW_h^K. \tag{18}$$

$$V_h = IW_h^V. \tag{19}$$

The introduction of the W-MSA module can significantly reduce the amount of computation. During the self-attention calculation, the MSA module must perform a correlation calculation for every two pixels in the feature map. However, when using the W-MSA module, we only need to divide the feature map into many windows according to the size of $M \times M$ and then perform the self-attention calculation on each window separately. It can be seen from Formula (20) and Formula (21) that the SwinT module can effectively reduce the amount of calculation compared with the ordinary transformer module, thus improving the real-time performance of the algorithm.

$$\Omega(\text{MSA}) = 4hwc^2 + 2(hw)^2 c, \tag{20}$$

$$\Omega(\text{W-MSA}) = 4hwc^2 + 2M^2 hwc, \tag{21}$$

where $h$, $w$, and $c$ represent the image's height, width, and length, respectively, and $M^2$ represents the number of windows.

### 2.4. SK convolution module

The human visual cortex will dynamically adjust the receptive field of neurons according to the different stimulus levels of targets of different sizes. Inspired by this, Li et al. [30] proposed a Selective Kernel Network (SKNet) model, which combines the idea of group convolution [31], dilated convolution [32], and the Squeeze-and-Excitation Networks (SENet) [33], and realizes the addition of attention mechanism to convolution kernel of different sizes. The core structure of SKNet is the Selective Kernel(SK) convolution module, as shown in Fig. 6.

The SK convolution is mainly composed of three steps: split, fuse, and select. The split operation divides the input feature graph $X$ with the size of $h \times w \times C$ into two branches, using $3 \times 3$ convolution kernel and $5 \times 5$ convolution kernel respectively to obtain $\tilde{U}$ and $\hat{U}$. In order to enhance the ability of feature extraction, $3 \times 3$ dilated convolution kernel with a dilation rate of 2 is actually used instead of $5 \times 5$ convolution kernel [30]. The fuse operation is similar to the method used by SENet. First, $\tilde{U}$ and $\hat{U}$ are fused by element-wise summation to obtain $U$, then the global average pooling $F_{gp}$ is used to obtain the eigenvector $s$ of $1 \times 1 \times C$, and then the full connection $F_{fc}$ is used to reduce the dimension to obtain the eigenvector $z$ of $1 \times 1 \times d$, and finally, the two Softmax functions are used to output the matrices $a$ and $b$ containing the weight information of different branch channels with the size of $1 \times 1 \times C$. The select operation weights the original characteristic graphs $\tilde{U}$ and $\hat{U}$ with the weight matrices $a$ and $b$, respectively, and then adds them to obtain the output characteristic graph $V$. The element c of s can be expressed as:

$$s_c = F_{gp}\left(U_c\right) = \frac{1}{H \cdot D} \sum_{i=1}^{H} \sum_{j=1}^{D} U_c(i,j), \tag{22}$$

where $H$ is the height of the feature map, $D$ is the width of the feature map, $c$ represents the number of channels, and $F_{gp}$ is the global average pooling function.

The compact feature $z$ that can be converted into precise and adaptive selection weights can be expressed as:

$$z = F_{fc}(s) = \delta\left(H\left(D_s\right)\right), \tag{23}$$

where $F_{fc}$ is the fully connection layer function, $\delta$ is the ReLU function, $H$ is the BatchNormalization function, and $D$ is $d \times c$ dimension, $D \in R^{d \times c}$. $d$ is expressed as $d = max(c/r, L)$, where $r$ is the reduction ratio and $L$ is the minimum value of $d$.

In the select operation, the softmax function is used to calculate the weight of each channel of different receptive field information, weight $a_c$ and weight $b_c$ can be expressed as:

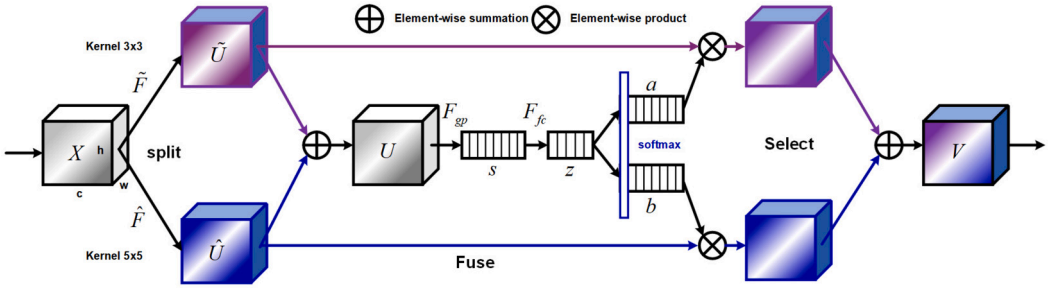



**Fig. 6.** The structure of the SK convolution.

$$a_c = \frac{e^{V_c z}}{e^{V_c z} + e^{H_c z}}, \tag{24}$$

$$b_c = \frac{e^{H_c z}}{e^{V_c z} + e^{H_c z}}, \tag{25}$$

where $V_c$ and $H_c$ are the matrices obtained when calculating $\tilde{U}$ and $\hat{U}$.

Finally, the extracted receptive field features of each scale are weighted and summed to obtain the feature map $V$; the characteristics of each channel of $V$ can be expressed as follows:

$$V_c = a_c \times \widetilde{U} + b_c \times \hat{U} \left(a_c + b_c = 1\right). \tag{26}$$

The SK convolution is no longer limited to the attention mechanism at channel or spatial but implements the attention mechanism for convolution kernels of different sizes so that the network can adaptively adjust its own structure. At the same time, the SK convolution is also a real-time plug-and-play module, which improves the accuracy of CV tasks such as semantic segmentation or target detection without adding too much computation to the network and is suitable for applications requiring high real-time performance, such as autonomous driving [34].

### 2.5. Proposed method

#### 2.5.1. ET-ARM module

The lanes often have some areas that are difficult to be identified. The semantic features of this area depend on the semantic feature information corresponding to some positions in the image [35]. Therefore, it is necessary to increase the perception field of the feature map to obtain more global context information. The disadvantage of the CNN in the context embedding block in capturing long-distance dependencies limits the further improvement of the segmentation model performance, and the global average pooling operation loses a lot of semantic features.

In order to solve the above problems, by applying the SwinT module and channel attention mechanism to the road scene semantic segmentation model, this study proposes the ET-ARM model (see Fig. 7). First, by using the SK convolution module, the useful information is reasonably weighted, the information is selectively emphasized, and irrelevant, redundant information is suppressed to improve the segmentation accuracy of the network model. Then, the SwinT module calculates the correlation between each element and other elements in the feature map to obtain the weighted feature representation. At the same time, the residual structure is used to combine the input image's features with the weighted feature representation output by the SwinT module to obtain enhanced features so as to retain the spatial information of the features. Finally, the feature map is given weight on both the channel and spatial at the same time to make up for the disadvantage of the Context Embedding Block, which does not pay enough attention to important information features. At the same time, in order to reduce the number of model parameters to a certain extent, we use $3 \times 1$ asymmetrical convolution and $1 \times 3$ asymmetric convolution to replace a $3 \times 3$ standard convolution and asymmetric convolution can ensure the equivalence of the results with the conventional convolution while reducing the computational complexity.

Conventional convolution is limited by fixed geometric structures, and it is difficult to perceive the geometric deformation of lanes. The available information is limited, which easily leads to the loss of key feature information of the target. In order to improve the modeling ability of the convolution neural network for lanes, deformable convolution is introduced to replace ordinary convolution, and a deformable residual block is constructed. The sampling with offset is used to replace the fixed position sampling. The sampling position is automatically shifted to the position of the lane area so that the shape of the convolution kernel can adapt to the shape of the lane. The convolution layer of the ARM module is reconstructed to capture the micro-structural details of the lane in various shapes and sizes.

#### 2.5.2. U-FFM module

In the SwinT module, in order to achieve the global interaction of context semantic information as much as possible under the computational complexity that is linearly related to the image size, the shifted windows are used in the hierarchical representation to achieve the information interaction between different windows. In fact, when half a window is offset each time, there is still a large amount of context information that cannot be communicated well in space.

The low-level and high-level semantic feature information in the STDC method is fused by the FFM model. High-level semantic feature information plays a key role in improving the segmentation performance of lanes in the image. However, due to the large receptive field of the high-level semantic feature map, the spatial information of lanes will be implicitly lost, which leads to the lack of location information of lanes in the feature map generated by the FFM model, which is not conducive to the improvement of segmentation accuracy. If the high-level semantic feature information is directly upsampled, the edge, detail, and texture information of lanes will be lost. Therefore, the integration of multi-level and multi-scale feature information in this study can make spatial information, and semantic features complement each other to achieve better segmentation results. However, if the feature information is directly fused and the differences between semantic features at different levels are ignored, it will lead to false segmentation. Therefore, this study proposes the U-FFM module, which aims to improve the ability to interact with information further and expand the size of the receptive field. The specific structure is shown in Fig. 7.

$$K = k + (k-1)(r-1), \tag{27}$$

where $K$ is the size of the actual convolution kernel, $r$ is the dilation rate, and $k$ is the size of the original convolution kernel.

In the CNN, the size of the convolution kernel determines the size of the convolutional receptive field, and corresponding to it, receptive fields of different sizes are suitable for identifying and segmenting targets of different sizes. Because the target changes in the road scene during the driving of autonomous driving vehicles have multi-angle and multi-scale characteristics, the design of receptive fields that can fuse multi-scale plays an important role in improving the recognition ability and segmentation accuracy. Inspired by the DeepLab series of algorithms [36,37], an enhanced perception module containing three

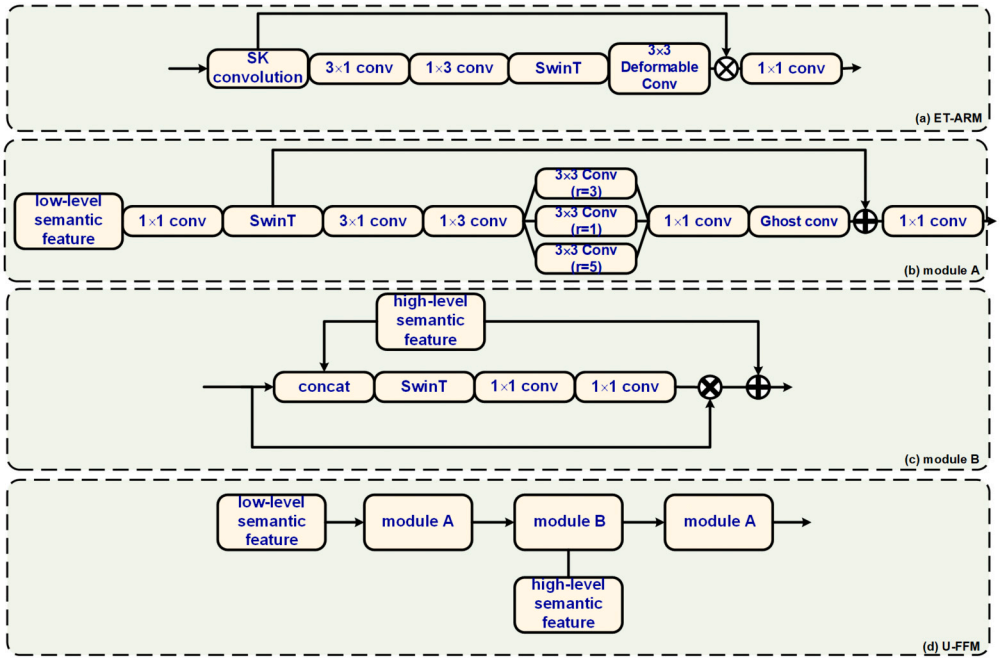



**Fig. 7.** The network structure of the ET-ARM module and the U-FFM module.

parallel branches was designed. In SwinT, the data stream is composed of vectors, and the data first passes through $1 \times 1$ convolution, which reduces the dimension of the length and width of the feature graph, thus reducing the network computing load and forming a multi-dimensional space feature map. Then, through the parallel dilated convolution [32] branches, the dilation rate of the three branches is 1, 3, and 5, respectively. According to Formula (27), the receptive fields with the size of $3 \times 3$, $7 \times 7$, and $11 \times 11$ are obtained, respectively. Because CNN, composed of many convolution modules, has a high degree of redundancy in the calculated intermediate feature map, it will increase computing costs. The use of dilated convolution to increase the size of the receptive field does not reduce the resolution of the feature map, and it can encode a wide range of contextual semantic information at different scales so that the feature map can obtain more accurate semantic and location information. When the scale and angle of lanes change constantly, it can effectively enhance the robustness of the lane recognition and segmentation methods. At the same time, this paper uses the lightweight module GhostNet to replace the partial convolution layer of the FFM model to speed up feature extraction. The detailed structure of UET-STDC is shown in the Table 4.

**Table 4**
For the detailed structure of UET-STDC, note that ConvX shown in the table refers to the Conv-BN-ReLU. KSize means kernel size. S, R, and C denote stride, repeat times, and output channels, respectively.

| Stage | Output size | KSize | S | STDC1 | | STDC2 | |
|---|---|---|---|---|---|---|---|
| | | | | R | C | R | C |
| Input | $224 \times 224$ | | | | 3 | | 3 |
| ConvX1 | $112 \times 112$ | $3 \times 3$ | 2 | 1 | 32 | 1 | 32 |
| ConvX2 | $56 \times 56$ | $3 \times 3$ | 2 | 1 | 64 | 1 | 64 |
| Stage3 | $28 \times 28$ | | 2 | 1 | 256 | 1 | 256 |
| | $28 \times 28$ | | 1 | 1 | 256 | 3 | 256 |
| Stage4 | $14 \times 14$ | | 2 | 1 | 512 | 1 | 512 |
| | $14 \times 14$ | | 1 | 1 | 512 | 2 | 512 |
| Stage5 | $7 \times 7$ | | 2 | 1 | 1024 | 1 | 1024 |
| | $7 \times 7$ | | 1 | 1 | 1024 | 1 | 1024 |
| ConvX6 | $7 \times 7$ | $1 \times 1$ | 1 | 1 | 1024 | 1 | 1024 |
| GlobalPool | $1 \times 1$ | $7 \times 7$ | | | | | |
| FC1 | | | | | 1024 | | 1024 |
| FC2 | | | | | 1000 | | 1000 |

### 2.5.3. The construction of self-made lane semantic segmentation dataset

The images of the self-made lane semantic segmentation dataset are mainly from the video frames collected by the vehicle camera. The scenes of video frames mainly include three types: the urban scene, the campus scene, and the expressway scene. The image obtained after image acquisition is a video image sequence in MPG format. First, convert the video image sequence from MPG format to AVI to ensure image clarity. After obtaining the video sequence in AVI format, convert the video sequence to a video frame for final annotation. Because the difference between adjacent video frames in the video sequence is small, and the number of samples in different categories may vary greatly, we manually selected some video frames from the collected video sequence as the final annotation pictures. The selection principle is to select pictures with large frame differences, large scene differences, and relatively complex scenes.

The semantic segmentation method assigns category labels to each pixel in the image, which requires the dataset to be used for fine-grained pixel-level annotation. We use the LabelMe annotation tool to carry out a pixel-level semantic annotation. The annotation category is divided into nine categories according to the lane category. By adding more images to the training set and expanding the sample space to enrich the training set, the data expansion approach may significantly reduce the overfitting of the semantic segmentation model and hence enhance its generalizability [38]. The UET-STDC method needs many images to train in real-world applications. The images captured by various vehicle-borne optical imaging sensors have varying image quality and styles, which may not meet the training requirements of the UET-STDC network model. However, collecting high-quality driving scene images in low-light environments is challenging, and the number of images in different categories may vary significantly. As a result, an efficient and appropriate method of data expansion is required.

In order to ensure the retention of the main contents of the road scene image, the specified transformation method is adopted in the data expansion stage. That is, the image is rotated, cut, and randomly cropped, and the image contrast is changed according to a certain probability. The 3000 original images used in the automatic driving environment semantic segmentation dataset assessed by the UET-STDC network model are

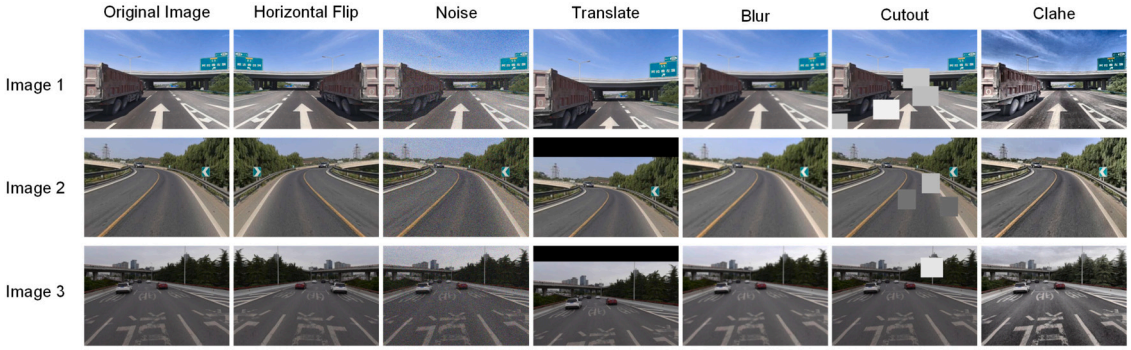



**Fig. 8.** The expansion effect of some image datasets.

**Table 5**
Experimental software version and hardware configuration.

| Name | Model and version |
|---|---|
| Integrated development environment | PyCharm Community Edition 2021.2.3 |
| GPU-accelerated library of primitives for deep neural networks | CUDA Deep Neural Network 8.4 |
| Graphics card | NVIDIA GeForce RTX 3070Ti |
| Memory capacity | 24 GB |
| Computer processor | AMD Ryzen 5 5600X 6-Core Processor |
| Computer vision and machine learning software library | OpenCV 4.6.0 |
| Deep learning software packages | PaddlePaddle 2.3.2 |

from videos we recorded while driving in Shanghai, Shenzhen, and other Chinese cities. Through rotation, clipping, random clipping, changing image contrast, and other technologies, the number of images in the dataset is increased to 6000 to improve the quality and diversity of existing image data. In Fig. 8, the expansion effects of several images are described. The semantic segmentation dataset of the automatic driving scene contains 5000 images, of which about 4500 are used as the training set, 1000 are used as the test set, and 500 are used as the verification set.

## 3. Semantic segmentation model training

In order to verify the practicability and effectiveness of the designed method, the TuSimple lane dataset and the self-made lane semantic segmentation dataset were used for training and test verification on a desktop computer equipped with NVIDIA RTX 3070Ti. The experimental environment uses the Windows 10 operating system, and the method proposed in this paper is built through the PaddlePaddle framework. In order to better compare with similar networks, the super parameters during our method training are mainly set according to the paper [38]. During network training, the input video frame size is set to $512 \times 1024$ and $768 \times 1536$, the batch size is set to 12, the optimizer selects Adam, the maximum iteration epoch is set to 200, the learning rate is 0.0025, and the momentum is 0.995. The software version and hardware configuration of this network model are shown in Table 5.

### 3.1. Evaluation indicators

This paper uses a variety of segmentation-related evaluation indicators to evaluate the performance of the semantic segmentation network from various aspects. The evaluation index of this experiment includes Precision, Recall, Intersection over Union (IoU), mean Intersection over Union (mIoU), kappa, Acc, and Dice. The formula of the Precision, formula, IoU, and mIoU coefficient can be expressed as follows:

$$Precision = \frac{p_{ii}}{p_{ii} + p_{ji}} \times 100\%, \tag{28}$$

$$Recall = \frac{p_{ii}}{p_{ii} + p_{ij}} \times 100\%, \tag{29}$$

$$IoU = \frac{p_{ii}}{\sum_{i=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ij} - p_{ii}}, \tag{30}$$

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{i=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ij} - p_{ii}}, \tag{31}$$

where $k$ is the number of categories of segmented objects, $i$ is the number of correctly segmented pixels, $p_{ii}$ is the number of pixels belonging to category $i$ which divided into category $i$, $p_{ij}$ is the number of pixels belonging to category $i$ but divided into category $j$, and $p_{ji}$ is the number of pixels belonging to category $j$ but divided into category $i$.

$$kappa = \frac{OA - p_c}{1 - p_c}, \tag{32}$$

where OA is the empirical probability of agreement on the label assigned to any sample, and pe is the expected agreement when both classifiers assign labels randomly. $p_c$ is estimated using a per-classifier empirical prior over the class labels.

The Acc coefficient is recognized as the most intuitive index in image segmentation evaluation. In short, Acc refers to the proportion of predicted correct pixel points in total pixels. The formula of the Acc coefficient can be expressed as follows:

$$Acc = \frac{\sum_{i=0}^{k} p_{ij}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}}, \tag{33}$$

where $k$ is the number of categories of segmented objects.

The formula of the Dice coefficient can be expressed as follows:

$$Dice = \frac{2\,|T \cap P|}{|T| + |P|}, \tag{34}$$

where $T$ represents the real target, $P$ represents the segmentation result. The value range of the Dice coefficient is $[0, 1]$; when the value is 1, it means that the real target is completely consistent with the segmentation result. In addition, False Positive (FP) represents the probability that pixels that are actually negative are predicted to be positive, while False Negative (FN) represents the probability that pixels that are actually positive are predicted to be negative.

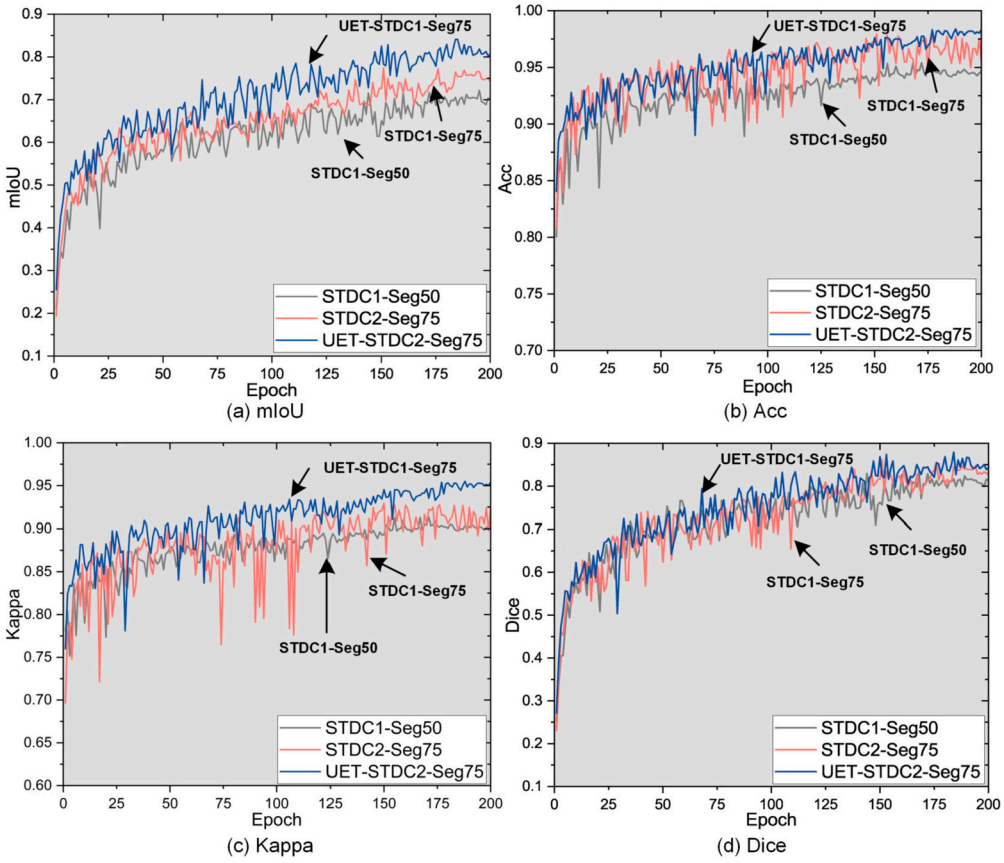



**Fig. 9.** Comparison of four evaluation index curves during trainings.

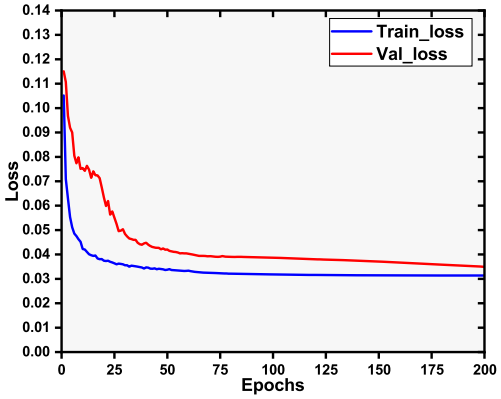



**Fig. 10.** The loss curve of UET-STDC1-Seg75. From this image, it can be seen that the UET-STDC1-Seg75 algorithm has a rapid reduction in loss value during the initial stage of the training process, and the loss curve gradually reaches a stable state after 50 epochs. At the same time, both training loss and validation loss decrease, and the two tend to stabilize with a small difference. This indicates that the algorithm performs well on the training set and can be well generalized to the validation set, meaning that the algorithm does not overfit or has a very small degree of overfitting.

### 3.2. Comparison with related methods

In order to further verify the segmentation performance of the UET-STDC method, the UET-STDC2-Seg75 method is compared with the related methods in terms of accuracy and speed. Table 6, Fig. 9, and Fig. 10 show the effect of region segmentation of lightweight series models and the comparison of the segmentation performance of network models. In order to improve the reliability of this research, the network model proposed in this study and the network model for comparison are deployed in the same experimental environment as Table 5 for comparison experiments. Table 6 shows that the mIoU of the UET-STDC2-Seg75 network model reaches 84.09%, which is 7.19% higher than the STDC2-Seg75 method, and 13.37% higher than the STDC method. The Acc indicator and Kappa indicator of the improved network model in this paper have reached 0.984 and 0.951, respectively, which are 1.17% and 2.91% higher than those of the STDC2-Seg75 method, and significantly improve the Dice indicator of the semantic segmentation, 3.73% higher than those of the STDC2-Seg75 method, however, due to the increase of computational load, the average segmentation speed of single image is slightly reduced by 6.43 FPS then the STDC2-Seg75 method. The experimental results show that the improved model performs well, and the segmentation accuracy is improved to a certain extent, which is more conducive to the deployment and application of the semantic segmentation model on vehicle-borne embedded terminals. In order to verify the effectiveness of the algorithm proposed in this study, we compared it with several state-of-the-art algorithms for detecting lane markings under low-light conditions in comparative experiments. From Table 6, it can be seen that the algorithm proposed in this study is superior to these state-of-the-art (SOTA) algorithms overall.

According to the segmentation results of the UET-STDC in Fig. 11, the contour of each category is relatively clear on the whole. In general, the UET-STDC method can effectively perform semantic segmentation in autonomous driving scenarios. However, although the STDC method can recognize the categories of some lanes, the segmentation boundary is fuzzy, and some double yellow lines are wrongly segmented. In summary, STDC has poor positioning accuracy for targets in autonomous driving scenes with sparse features, low visibility, and unclear targets, which easily leads to wrong segmentation. The improved method reduces the segmentation error rate of the autonomous driving scene and improves the segmentation accuracy. The reason is that the

**Table 6**
Comparison between mainstream semantic segmentation methods in self-made road scene semantic segmentation dataset. The experimental results show that the UET-STDC2-Seg75 model effectively balances the accuracy and real-time performance of semantic segmentation. Overall, the UET-STDC2-Seg75 model has better segmentation performance for lane markings under low-light conditions compared to other semantic segmentation algorithms.

| Model | Encoder | mIoU(%) | Acc | Kappa | Dice | FPS |
|---|---|---|---|---|---|---|
| STDC1-Seg50 | STDC1 | 70.59 | 0.954 | 0.906 | 0.826 | 181.36 |
| STDC2-Seg75 | STDC2 | 76.90 | 0.973 | 0.927 | 0.847 | 126.56 |
| LLFLD | ResNet-34 | 79.07 | 0.978 | 0.948 | 0.861 | 123.36 |
| LSTR | ResNet-18 | 74.54 | 0.971 | 0.934 | 0.852 | 114.18 |
| FastDraw | ResNet-50 | 72.33 | 0.926 | 0.925 | 0.843 | 98.50 |
| UET-STDC1-Seg50 | STDC1 | 77.42 | 0.976 | 0.935 | 0.853 | 177.29 |
| UET-STDC2-Seg75 | STDC2 | 84.09 | 0.984 | 0.953 | 0.879 | 120.13 |

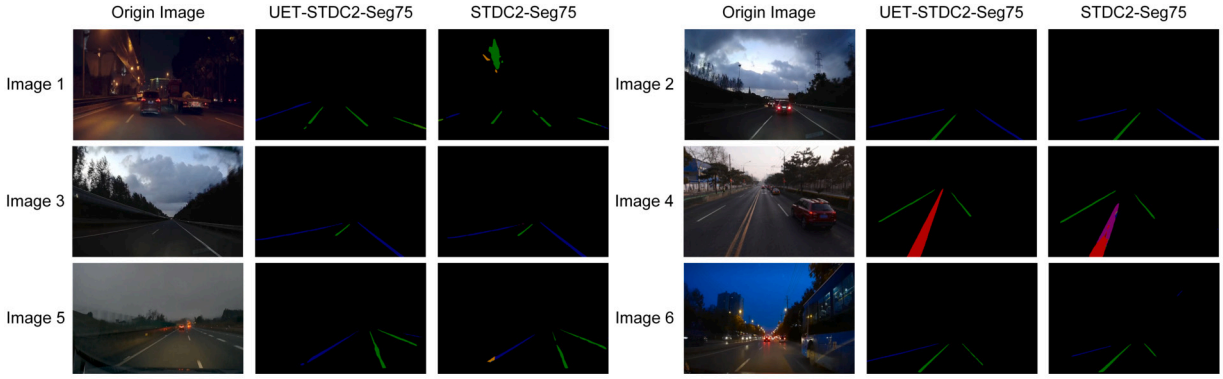



**Fig. 11.** Comparison of experimental results.

**Table 7**
Comparison with mainstream approaches on the Tusimple dataset.

| Model | Accuracy(%) | FP(%) | FN(%) | FPS(frame/s) |
|---|---|---|---|---|
| Blend Mask | 94.61 | 3.8 | 7.2 | 35 |
| YOLACT | 95.36 | 2.4 | 7.8 | 94 |
| SCNN | 96.53 | 6.1 | 1.8 | 24 |
| LaneNet | 96.38 | 7.8 | 2.4 | 57 |
| E2E_LMD | 96.04 | 3.1 | 4.1 | 63 |
| BiSeNetv2 | 95.89 | 5.2 | 3.4 | 98 |
| STDC2-Seg75 | 94.44 | 3.5 | 4.3 | 93 |
| **UET-STDC2-Seg75** | **98.53** | **2.1** | **2.5** | **91** |

Zero-DCE++ method is introduced to preprocess the captured image, thus enhancing the visibility of the image in the low-light environment. Meanwhile, the UET-STDC method improves the feature extraction module, which can extract more comprehensive semantic features of the autonomous driving scene to segment the targets in the autonomous driving scene accurately. At the same time, in order to verify the generalization performance of the algorithm proposed in this study, we validated the lane detection performance of the algorithm in scenes under normal illumination conditions. Fig. 12 visualizes the test results of the UET-STDC2-Seg75 algorithm in actual road scenarios, with solid lines representing the detected lane. The detection results indicate that the lane detection algorithm can accurately detect lanes.

The trained model was used to conduct experiments on the Tusimple dataset, and the lane detection results are shown in Table 7. Table 7 compares and analyzes the quantitative detection results based on STDC2-Seg75, UET-STDC2-Seg75, and several SOTA lane detection algorithms. It can be seen that both algorithms can effectively detect lanes and achieve real-time operation speed. The image frame processing speeds of UET-STDC2-Seg75 is 91 FPS, respectively. The UET-STDC2-Seg75 algorithm has higher detection accuracy, with a detection accuracy of 98.53%, which is 4.09% higher than the baseline algorithm. This indicates that the UET-STDC2-Seg75 algorithm is overall superior to the SOTA algorithm in terms of detection accuracy and speed.

## 4. Discussion

### 4.1. Ablation study

By comparing Model A and Model B in Table 8, it can be seen that the introduction of the SK convolution module in UET-STDC1-Seg50 increases mIoU by 0.78% and decreases FPS by 0.28. By comparing Model A and Model C, it can be seen that the introduction of the attention mechanism in the UET-STDC1-Seg50 algorithm will increase the test time slightly, but the mIoU will be significantly improved. The comparison between Model C and Model D shows that the parallel branches can significantly improve the mIoU of the UET-STDC1-Seg50 algorithm, but the running time of the algorithm is slightly increased. By comparing Model D and Model E, it can be seen that the introduction of the Ghostconv in the UET-STDC1-Seg50 algorithm will increase the mIoU slightly, but the test time will be slightly increased. It can be verified by comprehensive analysis in Table 8 that the modules mentioned have played a certain role in improving the accuracy of the UET-STDC1-Seg50 algorithm.

The above analysis results show that the illumination in the autonomous driving scene mainly comes from natural light, and the driving scene often changes. Therefore, there may be low illumination in the driving scene, which reduces the ability of the network to extract features. In this paper, we use the Zero-DCE++ algorithm to improve the problem of feature extraction ability decline caused by low-light levels through image enhancement and obtain good results. In addition, the camera moves with the vehicle, accompanied by the jitter, so some images are blurred, and the image background changes quickly. The SwinT and SK convolution modules are introduced in this method to enhance the ability to extract object contour and texture features. In addition,

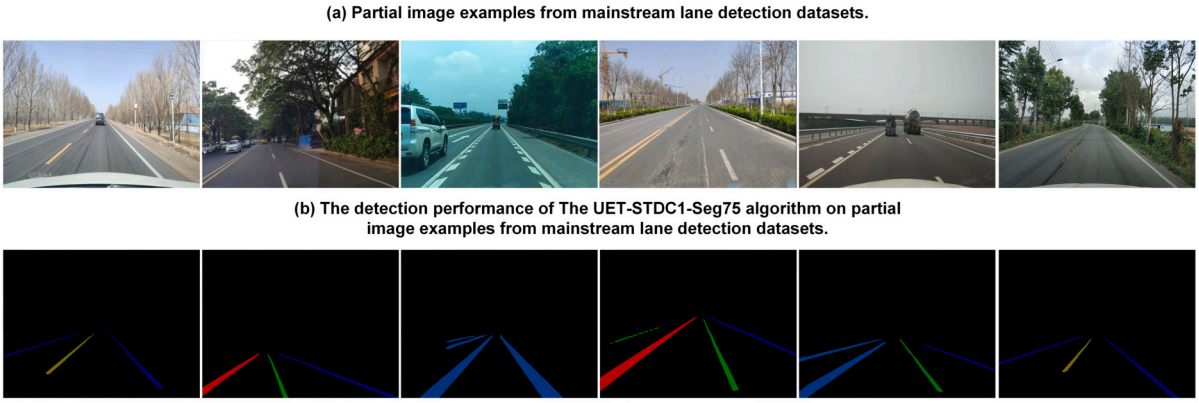



**Fig. 12.** Examples of detection results in complex background images of the ApolloScape dataset [39] and the CurveLanes dataset [40].

**Table 8**
Comparison between related segmentation algorithms.

| Model | ET-ARM | U-FFM module | | | mIoU(%) | FPS |
|---|---|---|---|---|---|---|
| | SK convolution | SwinT | Ghostconv | Parallel branches | | |
| Model A | × | × | × | × | 70.59 | 181.36 |
| Model B | ✓ | × | × | × | 71.37 | 181.08 |
| Model C | ✓ | ✓ | × | × | 75.62 | 178.93 |
| Model D | ✓ | ✓ | × | ✓ | 77.13 | 177.45 |
| Model E | ✓ | ✓ | ✓ | ✓ | 77.42 | 177.29 |

the shape and size of the scene change with the movement of vehicles and the switching of view angles. A new connection mechanism is introduced to enhance the semantic segmentation model's feature extraction and representation ability. The above comparison further verifies the advantages of this method.

### *4.2. Application in autonomous driving*

In the application of ADS, the DRL method is widely used by researchers [41,42], such as the method based on the actor-critic framework (see Fig. 13). This method can make driving behavior decisions according to the driving environment. However, due to the high training requirements of the DRL method, the depth of the DRL method's environment perception neural network cannot be designed too deep, which leads to the limited driving environment state that the DRL method can perceive in a low-light environment and usually cannot make driving behavior decisions. At present, the semantic segmentation method has been widely used as an auxiliary RL task. The proposed UET-STDC semantic segmentation method can quickly segment the driving area, isolation fence, roadside buildings, vegetation, vehicles, and pedestrians in the autonomous driving environment in a low-light environment, mark them in video frames, and transmit the video frames to the DRL method as states. The DRL method can output the corresponding autonomous driving decisions according to the video frames output by the UET-STDC semantic segmentation method.

Currently, most autonomous driving vehicles can only make decisions, plan, and control in relatively closed scenes with high-precision maps. These methods are based on preset rules. When the scene perceived by the vehicle is the same as the set scene, the behavior is determined according to the defined rules. Due to the complexity of autonomous driving scenes, preset rules often cannot cover all scenes, leading to the control failure of the autonomous driving vehicle, even causing fatal consequences. The vehicle's autonomous learning capability, driven by data and artificial intelligence algorithms, is expected to cope with complex autonomous driving scenarios. Autonomous vehicles need to perceive the complex environment to make decisions and control the vehicle to reach the next environmental state. Constant interaction with the environment is needed to achieve autonomous learning and constantly optimize the decision-making control scheme. This process conforms to the working paradigm of Reinforcement Learning (RL), which models continuous decision-making problems as Markov processes and finds the optimal solution by solving the Bellman equation. However, due to the high computational complexity of RL, it is unable to solve the problem of continuous state space and behavior space. DL has a strong perception ability and nonlinear function fitting ability, which makes it possible to solve continuous state space problems. The autonomous driving system is mainly composed of perception, planning, decision-making, control, and other modules, while RL is good at dealing with sequential decision-making problems with the goal of maximizing cumulative returns. Researchers have begun using DRL as an end-to-end solution for automatic driving, and the strategy directly outputs control signals such as vehicle throttle and steering. In addition, there is also relevant research on autonomous driving behavior decision-making based on DRL. In the future, the work that can be improved in this study includes:

1. Compared with the actual scenes, the types of lanes collected in the self-made lane semantic segmentation dataset are relatively small. In future research, we will study the application of transfer learning in DRL. Transfer learning can transfer the knowledge or skills learned in a single scene to other environments and the knowledge learned in the virtual environment to the real scene, which will undoubtedly greatly promote the application of driving strategies based on DRL in reality.
2. The training of the traditional DRL uses a random strategy for initialization, so the efficiency of agent exploration is relatively low, and the training process is long. In the following research, we will use imitation learning to obtain an initialization strategy based on expert data, thus greatly accelerating the training process.
3. The reward function is the key factor in determining whether the agent can learn effective strategies. In traditional DRL methods, the design of the reward function needs to adjust the coefficients of different reward items in the reward function, which is highly subjective. In future research, we consider using the Inverse reinforcement learning (IRL) method to avoid the process of design-

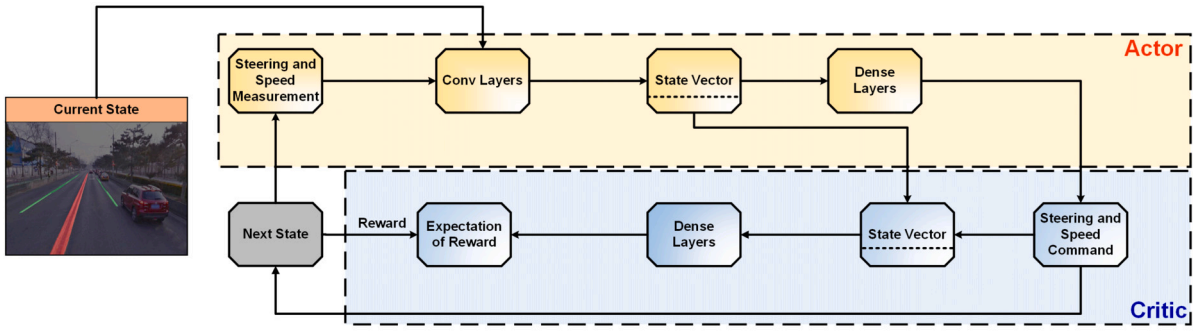



**Fig. 13.** The architecture of the Actor-Critic frame.

ing reward functions and automatically find more effective reward functions.

## 5. Conclusions

In this study, aiming at the problem of road scenes semantic segmentation under low-light conditions, where objects have many categories, and the same object has many different forms and occludes each other, the UET-STDC semantic segmentation method is proposed. The UET-STDC semantic segmentation method is applied to the semantic segmentation of road scenes and has achieved good results, which can meet the practicality and real-time of the semantic segmentation method under low-light conditions. The experimental results show that the Dice coefficient, the recall rate, and the accuracy rate of the UET-STDC semantic segmentation method reach 97.4%, 98.3%, and 96.5%, respectively, and the segmentation speed reaches 64 FPS. The overall performance is better than the current mainstream semantic segmentation network. In the future, we will combine research on migration learning and lane detection to increase the comprehensiveness and generalization of semantic segmentation and improve the efficiency of semantic segmentation models so as to improve the level of unmanned driving technology and further ensure the safety of autonomous driving. At the same time, the lightweight semantic segmentation method proposed in this study can provide annotated video frames for the DRL method, thus providing assistance in the decision-making process. We will further optimize the UET-STDC method so as to provide a new feasible technical solution for the mature application of the DRL method in the field of autonomous driving.

On the other hand, although the UET-STDC method proposed in this paper has made some progress, there are still some shortcomings and limitations that need to be further improved in the following aspects:

1. Although the method proposed in this paper based on UET-STDC can achieve real-time segmentation of lane images, the segmentation accuracy is far behind some complex networks, and further optimization is needed in data processing, model design, and training methods, such as introducing more image preprocessing and other data enhancement strategies and considering the use of pre-trained backbone networks in model design.
2. In this paper, the STDC backbone model is used to build a lightweight network. Although it reduces the size of the feature map space flowing in the network, it also leads to a slightly bloated convolution layer. At present, there are many other solutions in the lightweight network model, such as knowledge distillation, which can be further applied in this study.
3. The method of video data processing in this paper is still to split the video into a single image for frame-by-frame processing. If we can learn the spatial domain information in video data through weakly supervised learning and unsupervised learning, it will help to improve the stability of video segmentation results.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgments

This work was supported in part by the Joint Project of Nature Science Foundation of Liaoning Province of China (Program SN: 2021-KF-11-02).

## References

[1] H. Gajjar, S. Sanyal, M. Shah, A comprehensive study on lane detecting autonomous car using computer vision, Expert Syst. Appl. 233 (2023) 120929.
[2] C. Ke, Z. Xu, J. Zhang, D. Zhang, Combining low-light scene enhancement for fast and accurate lane detection, Sensors 23 (2023) 4917.
[3] Y. Wang, J. Zhang, Y. Chen, H. Yuan, C. Wu, Automatic learning-based data optimization method for autonomous driving, Digit. Signal Process. (2024) 104428.
[4] C. Xu, H. Liu, Q. Li, Y. Su, Driver's visual fixation attention prediction in dynamic scenes using hybrid neural networks, Digit. Signal Process. 142 (2023) 104217.
[5] H. Zhu, K.-V. Yuen, L. Mihaylova, H. Leung, Overview of environment perception for intelligent vehicles, IEEE Trans. Intell. Transp. Syst. 18 (2017) 2584–2601.
[6] H.-Y. Cheng, B.-S. Jeng, P.-T. Tseng, K.-C. Fan, Lane detection with moving vehicles in the traffic scenes, IEEE Trans. Intell. Transp. Syst. 7 (2006) 571–582.
[7] J. König, M.D. Jenkins, M. Mannion, P. Barrie, G. Morison, Optimized deep encoder-decoder methods for crack segmentation, Digit. Signal Process. 108 (2021) 102907.
[8] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, X. Wei, Rethinking bisenet for real-time semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9716–9725.
[9] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, Bisenet: bilateral segmentation network for real-time semantic segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 325–341.
[10] P. Yadav, N. Gupta, P.K. Sharma, Robust weapon detection in dark environments using yolov7-darkvision, Digit. Signal Process. 145 (2024) 104342.
[11] L. Tao, C. Zhu, G. Xiang, Y. Li, H. Jia, X. Xie, Llcnn: A convolutional neural network for low-light image enhancement, in: 2017 IEEE Visual Communications and Image Processing (VCIP), IEEE, 2017, pp. 1–4.
[12] C. Li, J. Guo, F. Porikli, Y. Pang, Lightennet: A convolutional neural network for weakly illuminated image enhancement, Pattern Recognit. Lett. 104 (2018) 15–22.
[13] C. Chen, Q. Chen, J. Xu, V. Koltun, Learning to see in the dark, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3291–3300.
[14] L.-W. Wang, Z.-S. Liu, W.-C. Siu, D.P. Lun, Lightening network for low-light image enhancement, IEEE Trans. Image Process. 29 (2020) 7984–7996.
[15] L. Hu, M. Qin, F. Zhang, Z. Du, R. Liu, Rscnn: a cnn-based method to enhance low-light remote-sensing images, Remote Sens. 13 (2020) 62.
[16] J. Li, J. Li, F. Fang, F. Li, G. Zhang, Luminance-aware pyramid network for low-light image enhancement, IEEE Trans. Multimed. 23 (2020) 3153–3165.
[17] C. Li, C. Guo, C.C. Loy, Learning to enhance low-light image via zero-reference deep curve estimation, arXiv preprint, arXiv:2103.00860, 2021.
[18] A. Zhu, L. Zhang, Y. Shen, Y. Ma, S. Zhao, Y. Zhou, Zero-shot restoration of underexposed images via robust retinex decomposition, in: 2020 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2020, pp. 1–6.

[19] F. Lv, F. Lu, J. Wu, C. Lim, Mbllen: low-light image/video enhancement using cnns, in: BMVC, vol. 220, 2018, p. 4.
[20] K.G. Lore, A. Akintayo, S. Sarkar, Llnet: a deep autoencoder approach to natural low-light image enhancement, Pattern Recognit. 61 (2017) 650–662.
[21] L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, J. Ma, Msr-net: low-light image enhancement using deep convolutional network, arXiv preprint, arXiv:1711.02488, 2017.
[22] A. Mittal, R. Soundararajan, A.C. Bovik, Making a "completely blind" image quality analyzer, IEEE Signal Process. Lett. 20 (2012) 209–212.
[23] K. Sharifi, A. Leon-Garcia, Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video, IEEE Trans. Circuits Syst. Video Technol. 5 (1995) 52–56.
[24] C.M. Bishop, N.M. Nasrabadi, Pattern Recognition and Machine Learning, vol. 4, Springer, 2006.
[25] J. Liu, M. Xu, X. Xu, Y. Huang, Nonreference image quality evaluation algorithm based on wavelet convolutional neural network and information entropy, Entropy 21 (2019) 1070.
[26] Y. Fang, K. Ma, Z. Wang, W. Lin, Z. Fang, G. Zhai, No-reference quality assessment of contrast-distorted images based on natural scene statistics, IEEE Signal Process. Lett. 22 (2014) 838–842.
[27] H. Li, P. Xiong, H. Fan, J. Sun, Dfanet: deep feature aggregation for real-time semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9522–9531.
[28] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, N. Sang, Bisenet v2: bilateral network with guided aggregation for real-time semantic segmentation, Int. J. Comput. Vis. 129 (2021) 3051–3068.
[29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[30] X. Li, W. Wang, X. Hu, J. Yang, Selective kernel networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 510–519.
[31] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (2017) 84–90.
[32] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, arXiv preprint, arXiv:1511.07122, 2015.
[33] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[34] F. Chen, J. Wei, B. Xue, M. Zhang, Feature fusion and kernel selective in inception-v4 network, Appl. Soft Comput. 119 (2022) 108582.
[35] F. Yuan, K. Li, C. Wang, J. Shi, Y. Zhu, Fully extracting feature correlation between and within stages for semantic segmentation, Digit. Signal Process. (2022) 103578.
[36] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2017) 834–848.
[37] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818.
[38] J. Chang, S. Guan, Class highlight generative adversarial networks for strip steel defect classification, Int. J. Pattern Recognit. Artif. Intell. 36 (2022) 2252004.
[39] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, R. Yang, The apolloscape open dataset for autonomous driving and its application, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2019) 2702–2719.
[40] H. Xu, S. Wang, X. Cai, W. Zhang, X. Liang, Z. Li, Curvelane-nas: unifying lane-sensitive architecture search and adaptive point blending, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16, Springer, 2020, pp. 689–704.
[41] X. Liu, Y. Lu, X. Liu, S. Bai, S. Li, J. You, Wasserstein loss with alternative reinforcement learning for severity-aware semantic segmentation, IEEE Trans. Intell. Transp. Syst. 23 (2020) 587–596.
[42] P. Shang, X. Liu, C. Yu, G. Yan, Q. Xiang, X. Mi, A new ensemble deep graph reinforcement learning network for spatio-temporal traffic volume forecasting in a freeway network, Digit. Signal Process. 123 (2022) 103419.

**Yang Liu** was born in Liaoning, China, in 1993. He received the B.E. degrees in measurement and control technology and instrumentation program control from Shenyang Aerospace University, Shenyang, China, in 2016, He received the M.Sc. degrees in applied mathematics from Liaoning Technical University, Fuxin, China, in 2020, respectively, where he is currently pursuing the Ph.D. degree in Northeastern University, Shenyang, China. His current research interests include deep learning and reinforcement learning.

**Yongfu Wang** received the Ph.D. degree in control science and engineering from Northeastern University, Shenyang, China, in 2005. He is currently a Professor with the School of Mechanical Engineering and Automation, Northeastern University, Shenyang, China. His research interests include unmanned vehicles, intelligent modeling, and control of mechatronic systems.

**Qiansheng Li** is currently an Engineer with the Dalian Power Plant of HUANENG Power Inti Inc. He was with Dalian Power Plant of HUANENG Power Inti Inc for 23 years. His research interests include the modeling, optimization and control for complex industrial processes, intelligent control, and reinforcement learning.