A Semantic Segmentation Method for Road Scene Images Based on Improved DeeplabV3+ Network

Lihua Bi[1], Xiangfei Zhang[2], Shihao Li[3], Canlin Li[2]
School of Software Engineering, Zhengzhou University of Light Industry, Zhengzhou, China[1]
School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou, China[2]
School of Information Science and Technology, Beijing Forestry University, Beijing, China[3]

***Abstract*—Semantic segmentation of road scenes plays a crucial role in many fields such as autonomous driving, intelligent transportation systems and urban planning. Through the precise identification and segmentation of elements such as roads, pedestrians, vehicles, and traffic signs, the system can better understand the surrounding environment and make safe and effective decisions. However, the existing semantic segmentation technology still faces many challenges in the face of complex road scenes, such as lighting changes, weather effects, different viewing angles and the existence of occlusions. Combined with the actual road scene image, this paper improves DeeplabV3+ network and applies it to semantic segmentation of road scene image, and proposes a semantic segmentation method of road scene image based on improved DeeplabV3+ network. By adding enhancement strategies for road scene images and hyperparameter adjustment, the method improves the training process of DeeplabV3+ network, and uses SK attention mechanism to improve the feature fusion module in DeeplabV3+, so as to improve the segmentation effect of road scene images. After the validation of Cityscapes and other data sets, the segmentation accuracy index mIoU of the proposed method reaches 79.8%, which can predict better semantic style effect, effectively improve the segmentation performance and accuracy of the model, and achieve better segmentation index results in the comparison network, and the subjective visual effect of the segmentation is also better.**

***Keywords*—Image enhancement; attention mechanism; semantic segmentation; road scene images**

I. INTRODUCTION

Image semantic segmentation [1] aims to provide richer image semantic information for pixel-level image classification tasks. The need for semantic segmentation is crucial for applications in scenarios that require high-precision target segmentation, such as autonomous driving [2-4], environmental monitoring [5-7], augmented reality [8-10], and security surveillance [11-13]. By accurately segmenting objects in an image, more accurate scene analysis, target recognition and decision making can be achieved, thus enhancing system performance and application experience.

In recent years, the field of automatic driving is constantly developing, and the semantic segmentation technology of road scene plays an important role in the automatic driving system. Image semantic segmentation provides the autonomous driving system with rich road information and high-level understanding of the image, including the classification and accurate positioning of the target, so that the autonomous vehicle can fully understand the complex traffic situation around. However, there are still some specific problems in the existing road scene semantic segmentation technology, such as: insufficient robustness under different lighting conditions, which leads to the deviation of segmentation results. Due to the influence of bad weather (such as rain, snow, fog), the segmentation accuracy decreases significantly. And in crowded and dynamic scenes, it is easy to appear occlusion and confusion. These deficiencies limit the decision-making ability of autonomous vehicles in complex environments, and further research is needed to address these challenges to improve system safety and reliability.

With the continuous development of hardware level and computing power, the rapid development and application of deep learning technology provides new ideas for semantic segmentation research, and deep learning methods can significantly improve the accuracy of semantic segmentation. Full Convolutional Network (FCN) [14] is an important method to deal with image segmentation tasks using deep learning technology, which opens a new era of achieving high-precision semantic segmentation with deep learning as the core technology. However, FCN still has some shortcomings, for example, the relationship between pixels is not fully utilized and the results obtained are still not fine enough. Researchers have gone on to propose many network models with better segmentation results based on different technical features. Badrinarayanan et al. proposed SegNet [15] based on the structure of codec [16]. The innovation of SegNet network is that it reduces the number of fully-connected layers as well as the number of parameters and storage space of the streamlined model, and also outputs the indexing information in the pooling process to improve the image segmentation accuracy and efficiency of the decoding process. Noh et al. proposed DeconvNet [17], which improves the segmentation performance by introducing an inverse convolution layer at the decoder side to recover the resolution of the feature map through the coding-decoding structure. Olaf Ronneberger et al. proposed U-Net [18], whose symmetric coding and decoding network structure captures semantic information at different levels, pinpoints feature map information during up and down sampling, and preserves more information about the original image.

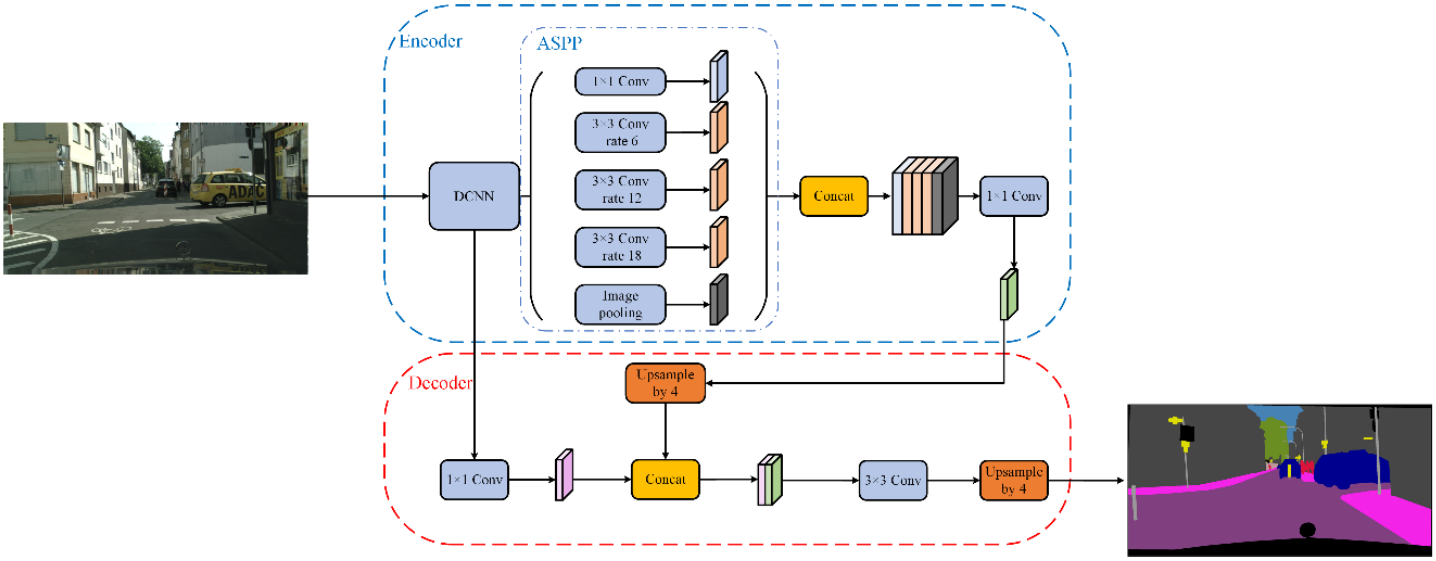


Fig. 1. DeeplabV3+ network architecture.

PSPNet [19] uses the ResNet [20] network as a backbone network, with modifications that use an additional auxiliary loss function, and two losses are assigned different weights, which promotes rapid convergence of the model. Overall, the traditional symmetric coding and decoding structure has more training parameters and complex structure, so it is less effective for the application of scenarios with real-time requirements.

The Deeplab network was proposed by the Google Brain team to solve the problems of category imbalance, voids, and edge details that are difficult to handle in semantic segmentation. Nowadays, Deeplab has developed a series of networks that are constantly employing new techniques and structural optimization algorithms to enhance their performance in various domains. DeepLabV1 [21] utilizes dilated convolutions to expand the receptive field and employs a fully connected conditional random field to enhance detail capturing capability, refining the segmentation object edges. DeepLabV2 [22] introduces the ASPP module on the basis of DeeplabV1, utilizing dilated convolutions with different dilation rates to extract feature information at different scales, enhancing the model's adaptability to objects of different scales, and also incorporating inverse convolutions and batch normalization techniques. DeeplabV3 [23] further expands the depth and width of the network on the basis of V2, adopts cascading or parallel mode to arrange cavity convolution with different cavity rates, optimizes ASPP module, and captures multi-scale features more effectively. DeeplabV3+ [24] extends DeepLabv3 by adopting an encoder-decoder structure to achieve better semantic segmentation performance. This paper proposes a semantic segmentation method for road scene images based on improved DeeplabV3+ network. The backbone network uses lightweight MobileNetV2, combined with the actual application of road scene image, to enhance image data and hyperparameter adjustment, aiming at improving the accuracy and efficiency of semantic segmentation of road scene image. We introduced SK attention module to optimize the feature fusion module in DeeplabV3+ to enhance the model's ability to capture key features. This improvement not only improves segmentation accuracy, especially in complex environments such as bright light, shadows, and dynamic scenes, but also helps reduce computational costs.

The main contributions of this paper are summarized as follows:

- This paper proposes a road scene image semantic segmentation method based on improved DeeplabV3+ network. To reduce the complexity of DeeplabV3+ network, lightweight MobileNetV2 is used as the backbone network.
- Increase the diversity of training data through image enhancement strategies, optimize the generalization of DeeplabV3+ network, and optimize the training process through hyperparameter adjustment.
- By introducing SK attention mechanism to optimize the feature fusion module in DeeplabV3+, adjust the feature weights and improve the accuracy of semantic segmentation.

The rest of this article is structured as follows: Section II describes the network structure of DeeplabV3+. In Section III, the semantic segmentation method of road scene image based on the improved DeeplabV3+ network is introduced in detail. Section IV describes the experimental setup of this paper and the experimental results with other methods. The final conclusion is given in Section V.

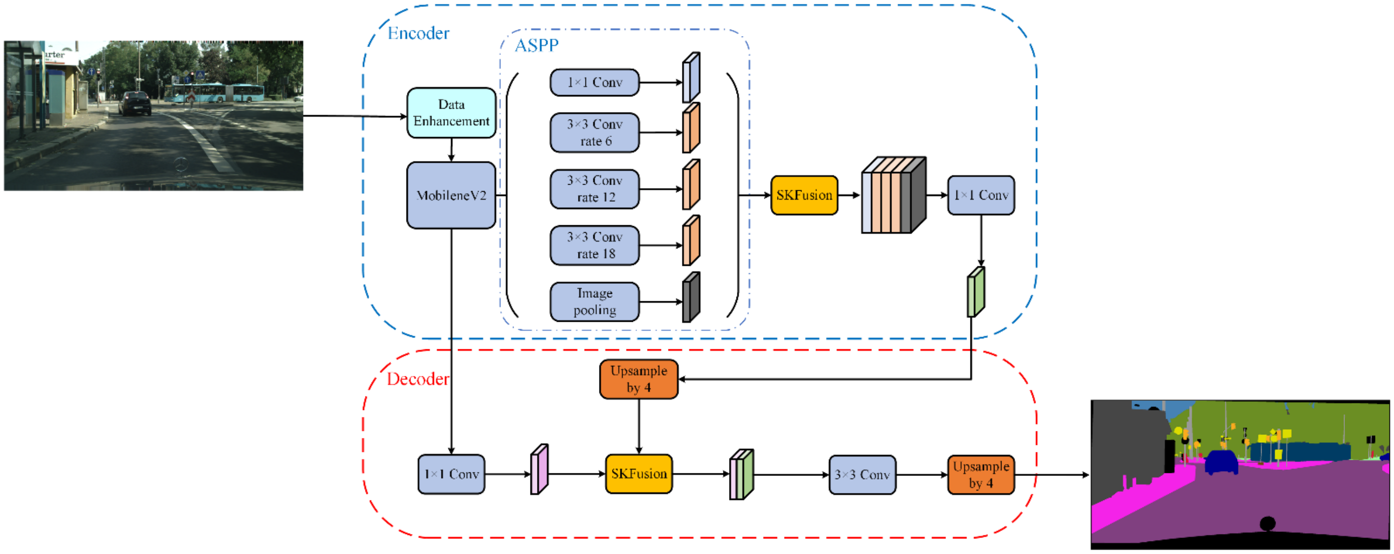


Fig. 2. Overall framework of road scene semantic segmentation based on improved DeeplabV3+.

II. DeepLabV3+ Network

The DeepLabv3+ network architecture is shown in Fig. 1. DeepLabv3+ uses an encoder-decoder structure to improve DeepLabv3. The encoder uses Atrous Spatial Pyramid Pooling (ASPP) to capture context information at different scales, while the decoder refines the target boundary to improve segmentation results. The ASPP module is used to capture semantic information of different scales, splice the feature information after multiple empty convolution operations with different sampling rates, and obtain the feature after 1×1 convolution. Another branch of the Deep Convolutional Neural Network (DCNN) uses 1×1 convolution to process the underlying features of the image to obtain the underlying features of the image. Then the feature is fused with the feature that has been subsampled four times, and the semantic segmentation prediction image is obtained after 3×3 convolution and four times subsampling.

III. A Semantic Segmentation Method for Road Scene Images Based on Improved DeepLabV3+ Network

A. *The Overall Framework of the Proposed Methodology*

In this paper, DeepLabv3+ model is used to improve image semantic segmentation in road scenes. In the semantic segmentation of road scene image based on DeepLabv3+ network, lightweight MobileNetV2 is adopted as the backbone in this paper. In the training process, the data is enriched by image enhancement. After passing through MobileNetV2 network, ASPP method is used to extract multi-scale feature information from images by different sampling rates. SK attention mechanism [25] is introduced to improve the feature fusion module, and feature fusion is performed on the feature mapping obtained by ASPP module to improve DeepLabv3+ network. The SK feature fusion module is also used in the later feature fusion of encoder and decoder. The improved Deeplabv3+ model is shown in Fig. 2.

B. *Image Enhancement*

Image enhancement refers to the process of generating new training samples through a series of transformation operations on the original image during image processing or pattern recognition tasks. Through image enhancement, the diversity of training data can be increased, which helps to improve the generalization ability and robustness of the model. The operation process of image enhancement in this paper is shown in Fig. 3.

1) Random left-right flip: In order to increase the diversity of data and make the model have better generalization ability and segmentation effect, the left and right flip of all the training set images is carried out with 50% probability. The same original image was randomly flipped left and right twice, and the resulting image was shown in Fig. 4.

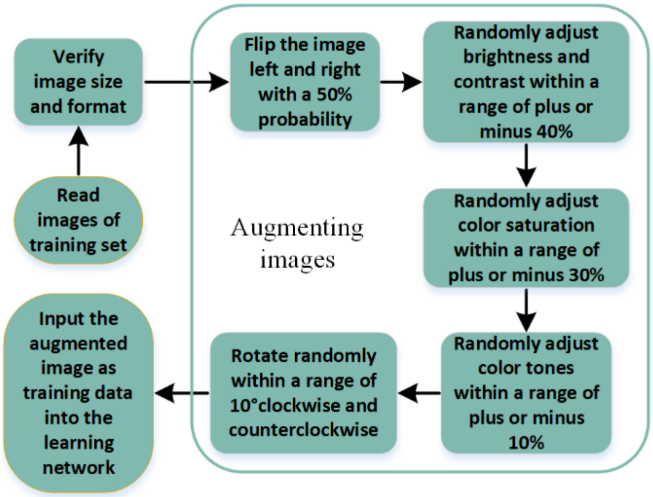
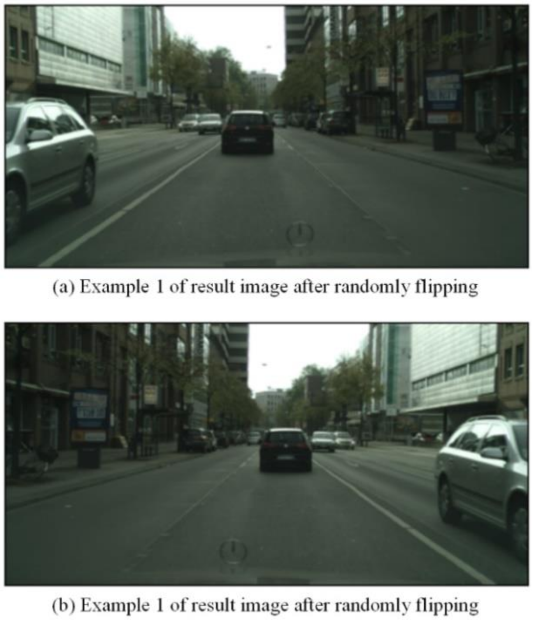

Verify image size and format
Flip the image left and right with a 50% probability
Randomly adjust brightness and contrast within a range of plus or minus 40%
Read images of training set
Augmenting images
Randomly adjust color saturation within a range of plus or minus 30%
Input the augmented image as training data into the learning network
Rotate randomly within a range of 10°clockwise and counterclockwise
Randomly adjust color tones within a range of plus or minus 10%

Fig. 3. Improved image augmentation operation process.

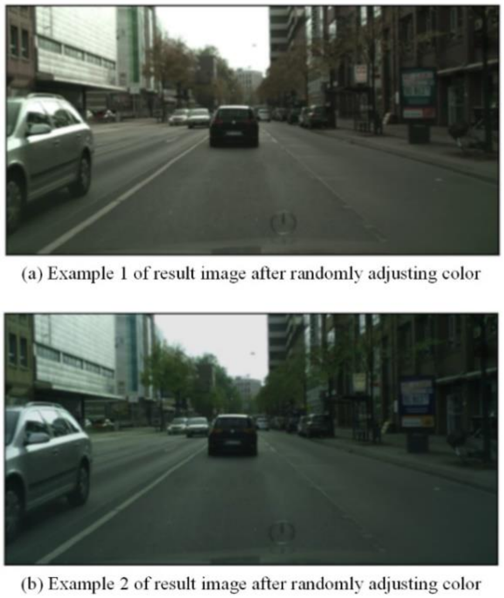


(a) Example 1 of result image after randomly flipping

(b) Example 1 of result image after randomly flipping

Fig. 4. Schematic diagram of random left-right flip.

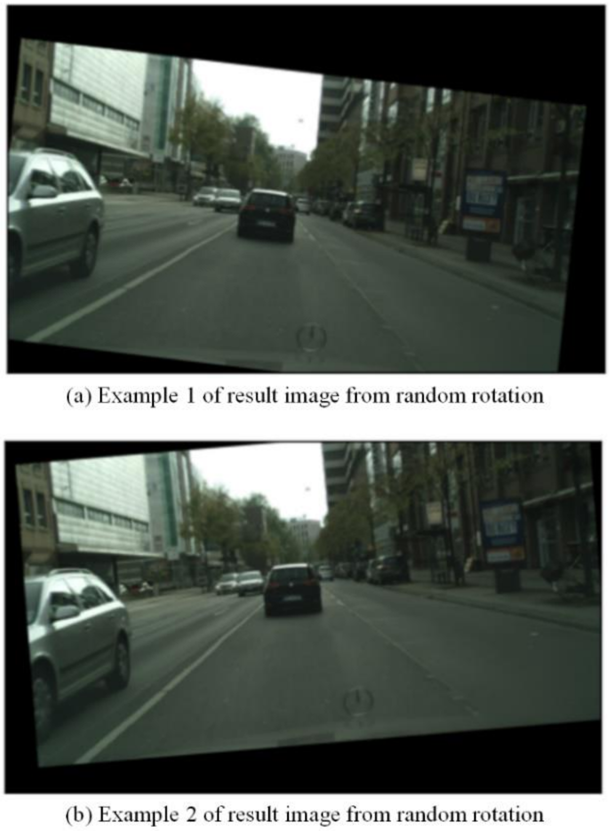
2) Random color adjustment: In the actual road scene image, may encounter a variety of different weather during driving, or due to the impact of camera shooting, resulting in a large difference in the brightness of the input image, so you need to adjust the image brightness, contrast, saturation and tone to increase the generalization ability of color. The pre-processed images are color-adjusted, the brightness and contrast are randomly adjusted within the range of plus or minus 40%, the color saturation is randomly adjusted within the range of plus or minus 30%, and the hue is randomly adjusted within the range of plus or minus 10%. Two random color adjustments were made to the same original training image, and the results were shown in Fig. 5.



(a) Example 1 of result image after randomly adjusting color

(b) Example 2 of result image after randomly adjusting color

Fig. 5. Schematic diagram of random color adjustment.



(a) Example 1 of result image from random rotation

(b) Example 2 of result image from random rotation

Fig. 6. Schematic diagram of random rotation angle.

3) Random rotation angle: In order to increase the generalization ability of oblique images, this paper added the processing of random rotation Angle, set the center of the image as the rotation center, and rotate the image randomly within the range of 10° clockwise and counter clockwise. The same picture was randomly rotated for two times, and the result was shown in Fig. 6.

C. SKFusion

When people observe things, they will selectively pay attention to the more important information, which is called attention. By continuously focusing on this key location to get more information and ignoring other useless information, this visual attention mechanism greatly improves the efficiency and accuracy of our processing on information. The attention mechanism in deep learning is similar to the attention mechanism in human vision, which is to focus attention on the important points with more information, select the key information, and ignore other unimportant information.

SKFusion in the improved DeepLabv3+ network is to adjust the weight of the feature map by using the SK attention mechanism after the concatenated features. The operation process is shown in Fig. 7. We first splice n features together by Concat operation, as shown in formula (1):

$$X = Concat(x_1, ..., x_n) \tag{1}$$

where, $(x_1, ..., x_n)$ a is n eigenvectors.

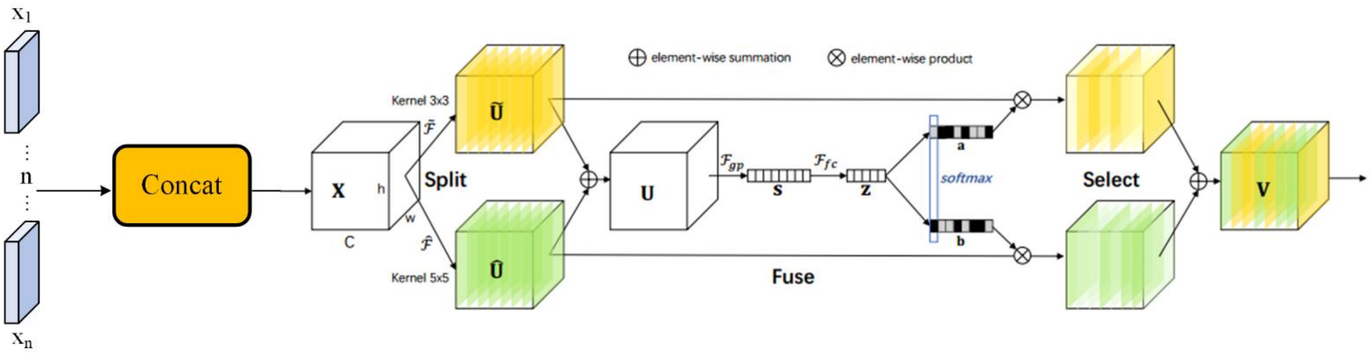


Fig. 7. SKFusion network architecture.

Then the obtained feature X is processed using the SK attention mechanism, and the processed feature is obtained, as shown in formula (2):

$$V = SKAttention(X) \tag{2}$$

SK attention mechanism [25] is mainly divided into three operations: Split, Fuse and Select. The Split operator produces multiple paths of different kernel sizes. The Fuse operators combine and aggregate information from multiple paths to obtain a global and comprehensive representation for selecting weights. The Select operator aggregates feature maps of cores of different sizes based on selection weights.

a) Split: Convolve the input feature graph X through a cavity of different receptive fields. Fig. 7 represents two groups of convolution operations, one with a convolution kernel of 3*3 to obtain the feature graph, and the other with a convolution kernel of 5*5 to obtain the convolution kernel, as shown in formulas (3) and (4):

$$\tilde{U} = Conv_{3*3}(X) \tag{3}$$

$$\hat{U} = Conv_{5*5}(X) \tag{4}$$

b) Fuse: To ensure that the information flow from multiple branches carries information of different sizes into the next layer of neurons, first fuse the results of multiple branches (two branches in Fig. 7) by summing the elements, as shown in formula (5):

$$U = \tilde{U} + \hat{U} \tag{5}$$

Then, the global average pooling of U is performed to obtain s, and the dimensionality is reduced by the fully connected layer to improve the efficiency and z is obtained, as shown in formulas (6) and (7):

$$s = F_{gp}(U) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U(i, j) \tag{6}$$

$$z = F_{fc}(s) \tag{7}$$

Where, F_{gp} represents the global average pooling operation, H×W is the spatial dimension size, and F_{fc} represents the fully connected layer.

c) Select: A new feature map computed from convolution kernels with different weights. First do softmax to calculate the weight of each convolution kernel, if there are two convolution cores, then a+b=1. Then, the final feature graph V is obtained by multiplying the weight elements of each convolution kernel, as shown in formula (8):

$$V = a \cdot \tilde{U} + b \cdot \hat{U}, a + b = 1 \tag{8}$$

D. Hyperparameter Adjustment

In the original training hyperparameters of the model, the batch size was set to 16. However, considering that there is still room for expansion in the memory of the graphics card, we have increased the batch size. Increasing the batch size has three following benefits. Firstly, the memory utilization is increased, and the parallelization efficiency of the large matrix multiplication is increased. Secondly, with the same amount of data, the number of iterations required to run through a training round is reduced, further increasing the processing speed. Thirdly, increasing the batch size within a certain range can make the determined descent direction more accurate, thus reducing training oscillations.

When the batch size increases to a certain extent, its determined descent direction may have been largely unchanged. However, since the accuracy of the final convergence is affected by a variety of factors, such as network structure, learning rate, etc., it does not mean that the larger batch size will necessarily get better results. In practice, when the batch size is increased to some extent, to achieve the optimal convergence accuracy, factors such as iteration need to be considered. So the adjustment strategy of our proposed method is to increase the batch size to 32.

For the epoch hyperparameter, the epoch was set to 30 in the initial experiment, which still has room for improvement. The complete process of running the model to complete one forward propagation and back propagation on all the data is called 1 training round, in other words, it means that all the training samples in the dataset have been trained once.

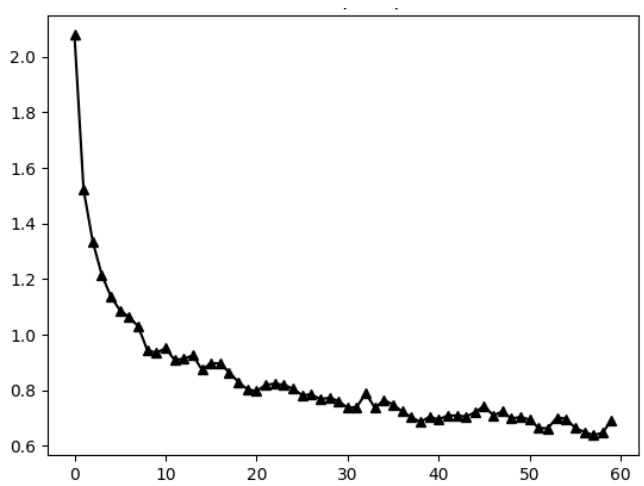


Fig. 8. Loss fluctuation per training set round.

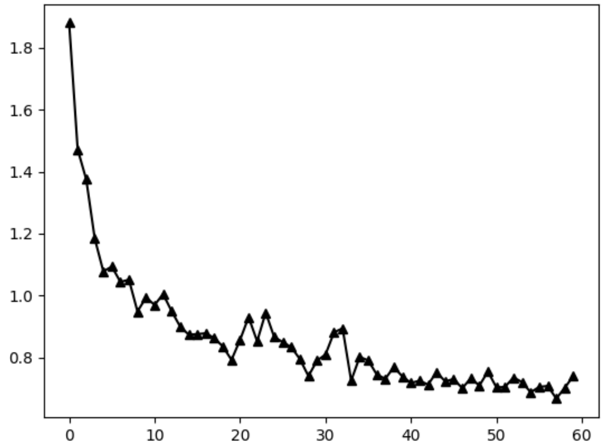


Fig. 9. Loss fluctuation per validation set round.

In the training process of gradient descent model, the neural network gradually transitions from the unfit state to the optimal fit state, then it will enter the overfitting state after reaching the optimal state. However, the number epoch of training rounds is not the larger the better, and it is generally set to between 20 and 200 to achieve good training results. The more diverse the data, the larger the corresponding training rounds. Our proposed method adjusts the epoch to 60.

We visualize the fluctuation of the loss function with the number of training rounds. Both the training and validation sets use the images contained in Cityscapes [26]. The loss fluctuations of training set are shown in Fig. 8, and the loss fluctuations of validation set are shown in Fig. 9.

Through recording the fluctuations of loss function during the training process of model, as can be seen from Fig. 8 and Fig. 9, as the number of training rounds increases, the loss decreases and the segmentation effect gradually improves. Due to the Adam optimization algorithm, after 40 rounds, the learning rate will gradually decrease after being adjusted by the Adam algorithm, so the weight change will also decrease, and the change of corresponding loss function will be minimal. After 60 rounds of training, a model with relatively good segmentation performance can already be obtained.

IV. Experimental Design and Analysis

A. Dataset

The rapid development of deep learning cannot be separated from the development of training data, and the preprocessed dataset can greatly facilitate the training process of image processing without having to put too much effort on the labeling of the dataset. In this paper, we choose the representative dataset Cityscapes [26] as the dataset for semantic segmentation of road scenes. Cityscapes is an image dataset for urban street scenes, and this dataset contains 5000 images that have been annotated at high quality pixel level and covers 19 categories with dense pixel annotations, among which 8 categories have instance-level segmentation. The dataset is divided into three parts: training, validation and testing with 2975, 500 and 1525 images, respectively. In addition, the dataset contains stereoscopic video sequences from street scenes in 50 different cities. The examples on raw and pixel-level labeled images of the Cityscapes dataset are shown in Fig. 10 and Fig. 11, respectively.

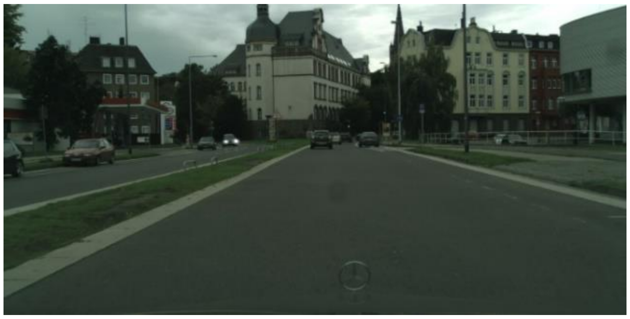


Fig. 10. Original image of the road scene.

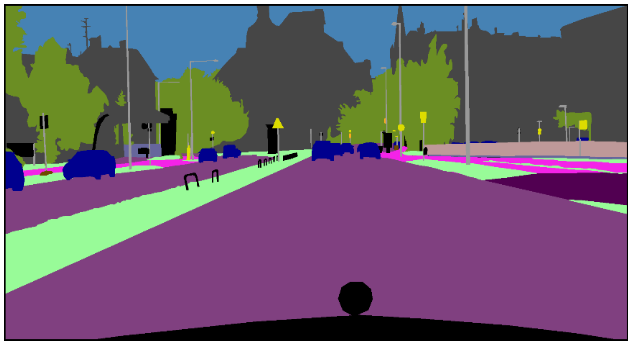


Fig. 11. Pixel-level labeled image.

Cityscapes is very large, with 20,000 weakly annotated frames in addition to 5,000 images with high-quality pixel-level annotations. In addition, the Cityscapes dataset provides fine and coarse metrics. The fine evaluation criterion is based on 5,000 images with fine labels, while the coarse evaluation criterion is based on 5,000 images with fine labels and 20,000 images with coarse labels.

B. Experimental Environment and Parameter Settings

Performing image processing using deep learning involves a large number of floating-point and matrix operations, and it has high requirements for the hardware and software environments, which can affect the effectiveness of deep learning model. The hardware environment for our experiments involves CPU processor Intel i7-9750H and 128G memory, as well as GPU processor NVIDIA GeForce GTX 1650 with 4G graphics memory. In terms of software environment, Windows 10 64-bit operating system is used, Pytorch deep learning framework is selected, and the classic general-purpose parallel computing architecture CUDA is applied. In addition, dependent libraries such as Numpy, OpenCV, and PIL are also used.

In order to ensure that the segmentation results are only affected by the model itself, the same hardware configuration and software parameter settings are used for different comparison models, so as to ensure the consistency of the experimental environments. For each comparison model, the experimental software and hardware environments are shown in Table I and Table II.

TABLE I. EXPERIMENTAL HARDWARE CONFIGURATION

Hardware	Configuration
CPU	Intel i7-9750H@2.60GHz
GPU	NVIDIA GeForce GTX 1650
Memory	128G
Video memory	4G

TABLE II. EXPERIMENTAL SOFTWARE CONFIGURATION

Software	Configuration
Operating system	Windows10 64-bit
Deep Learning Framework	Pytorch1.12
Programming language	Python3.8
Parallel computing architecture	CUDA 11.6
Main Dependency Libraries	Numpy、Matplotlib、Opencv

In this paper, the uniform parameter settings for model training are as follows. Cross-Entropy Loss (CE Loss) [27] with multiple classes is used. The optimizer adopts Stochastic Gradient Descent (SGD) strategy, where the learning momentum parameter is set to 0.9 and the weight decay is set to 0.00001. The initial learning rate LR is 0.0001, and the learning rate is dynamically adjusted by a poly strategy, which dynamically decreases with the increase of training iterations, and the current learning rate new_lr is updated as shown in formula (9).

$$new\_lr = LR * (1 - \frac{epo}{\max\_epo})^{power} \tag{9}$$

Where momentum power is 0.9, epo indicates the current number of training iteration, and the maximum number of training iteration max_epo is calculated as shown in formula (10).

$$\max\_epo = (\frac{M}{batchsize}) * epoch \tag{10}$$

Where M is the number of training samples 2975, the number epoch of training rounds is the total number of rounds that need to be trained, uniformly set to 60. According to the model size and the graphics memory, batchsize is set to 20. After the above parameter settings, all of comparison models can achieve good convergence results.

C. Comparison of Experimental Effect

To verify the effectiveness of the proposed method, several typical deep learning semantic segmentation networks were selected for experiments, including FCN, SegNet, DeeplabV3 and DeeplabV3+. By comparing with these networks, we aim to comprehensively evaluate the performance of our proposed methods in different scenarios. In the experiment, the segmentation effect is compared on the images of verification set as well as the campus images taken in the field to ensure the universality and reliability of the results. In addition, we will use multiple evaluation metrics such as mean crossover ratio (mIoU) and pixel accuracy (pixAcc) to provide a more comprehensive performance analysis.

1) Comparison of subjective effect: In this paper, five segmentation network models including FCN, SegNet, DeeplabV3, DeeplabV3+ and our improved DeeplabV3+ are applied to test and verify the semantic segmentation of the same image. Fig. 12 shows the segmentation effect diagram of some pictures. Columns (b), (c), (d), (e) and (f) respectively represent the segmentation effect corresponding to FCN, SegNet, DeeplabV3, DeeplabV3+ and our improved DeeplabV3+ network. Images 1 to 5 were selected from part of the validation set in the Cityscapes dataset, including several common road scenes. Images 6 and 7 were two road scenes in the university.

Compared with the segmentation effect of all images, it can be seen from Fig. 12 that SegNet has a slightly higher segmentation accuracy than FCN in general, but the segmentation effect needs to be improved and there are some problems of inaccurate segmentation of details. DeeplabV3 has a better segmentation effect than SegNet and can accurately predict the classification of some detailed images. Our improved DeeplabV3+ works best, with clearer categories. SegNet cannot clearly divide the shape of the category when dividing the categories such as cars and houses. FCN has some deviation when it does not distinguish pedestrian categories well, and the edges are fuzzy, while DeeplabV3 and DeeplabV3+ still have good segmentation effect. Since our improved DeeplabV3+ optimizes the training strategy through image enhancement and hyperparameter adjustment, and uses the SKFusion module DeeplabV3+ network structure, it is more sensitive to small targets and other information, and the segmentation effect is better.

It can be seen that the performance of the improved model at the junction of the motorway and the sidewalk has been significantly improved, and the jagged segmentation phenomenon of the original model has been successfully solved, making the boundary smoother and more natural. This improvement not only improves the visual effect, but also provides a more reliable basis for the division of pedestrians, vehicles and buildings in practical applications. In the segmentation results of images 6 and 7, we can clearly see that the improved DeeplabV3+ model shows higher accuracy and detail in processing various categories. Especially in the contours of pedestrians, cars and houses, the model can effectively distinguish the boundaries of complex backgrounds and reduce blurring. This further verifies the effectiveness and superiority of the model in multi-class segmentation tasks.

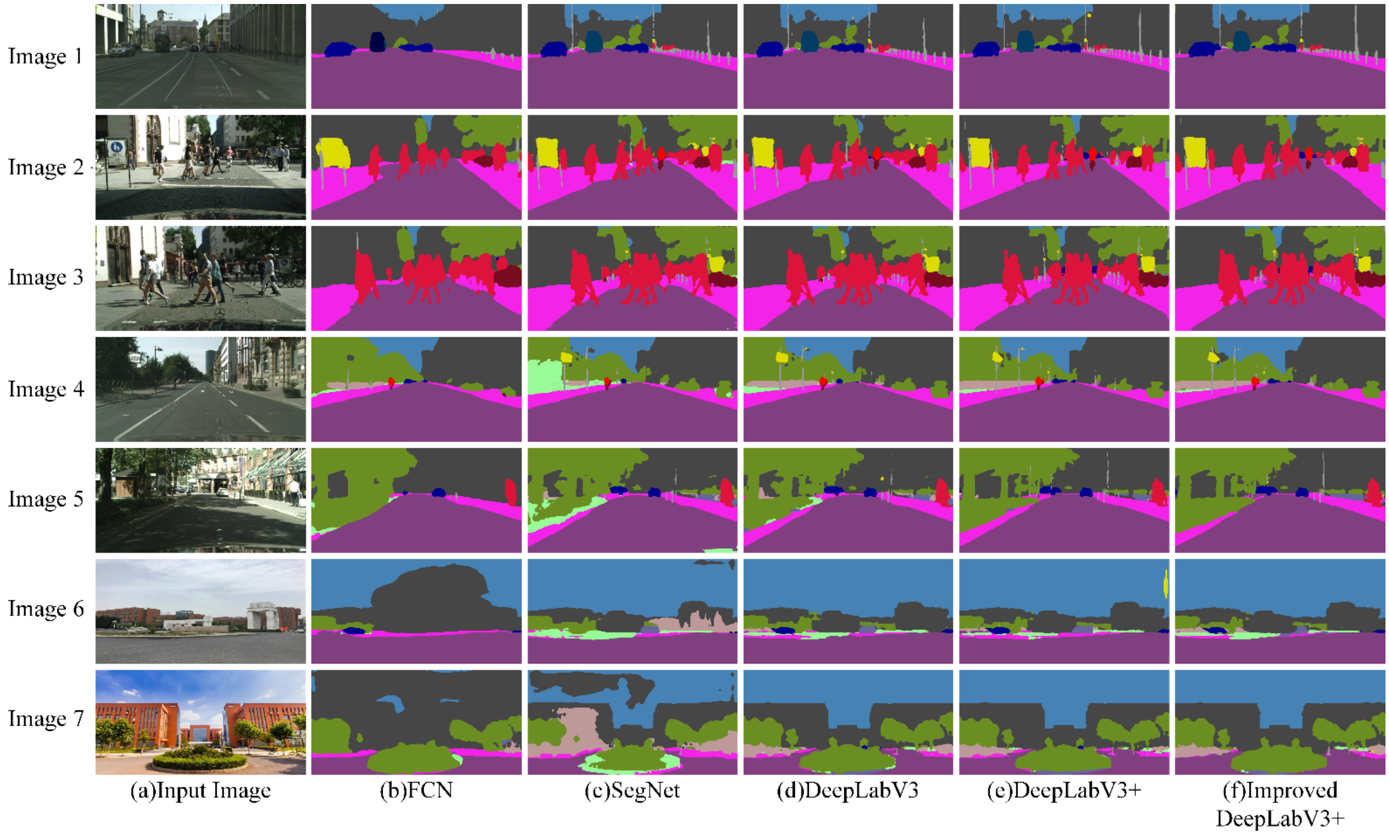


Fig. 12. Comparison of subjective effects. (a)Input Image; (b)FCN; (c)SegNet; (d)DeeplabV3; (e)DeeplabV3+; (f) Improved DeeplabV3+.

2) Comparison of objective effects: Regarding the objective performance evaluation metrics for semantic segmentation on road scene images, this paper adopts the commonly used evaluation metrics for semantic segmentation, pixAcc (PA) and mIoU. PA is the pixel accuracy rate, which is relatively simple to compute, and is the ratio of the number of correctly predicted pixels to the total predicted pixels. mIoU is a commonly used evaluation metric for semantic segmentation tasks, i.e., the average intersection and union ratio which is the ratio of the intersection and union of the true and predicted values. In semantic segmentation, the intersection and union ratio of a single category indicates the ratio of the intersection of the true and predicted values of the category to the union of the category, reflecting the classification accuracy of the model for each category and the overall segmentation effect.

For the test set contained in Cityscapes, the five models FCN, SegNet, DeeplabV3, DeeplabV3+, and our improved DeeplabV3+ are tested and the objective performance evaluation indicators are shown in Table III.

As a whole, combined with Fig. 12, our improved DeeplabV3+ model has better overall segmentation results, with finer segmentation results for some categories such as pedestrians, vehicles, traffic lights, and sign boards, as well as better segmentation results for more detailed categories such as travel lanes. After the validation on the Cityscapes validation set, the segmentation accuracy metric mIoU of our improved DeeplabV3+ reached 79.8%, achieving the best segmentation metric results.

TABLE III. COMPARISON OF OBJECTIVE PERFORMANCE EVALUATION INDICATORS OF FCN, SEGNET, DEEPLABV3, DEEPLABV3+, AND IMPROVED DEEPLABV3+ SEGMENTATION

Model	**pixAcc（%）**	**mIoU（%）**
FCN	81.8	32.5
SegNet	84.2	55.4
DeeplabV3	87.6	72.7
DeeplabV3+	89.4	76.5
improved DeeplabV3+	91.2	79.8

V. CONCLUSIONS

For the problem that the semantic segmentation effect of road scene image needs to be improved, this paper proposes a road scene image semantic segmentation method based on the improved DeeplabV3+ network by improving DeeplabV3+ network and applying it to the semantic segmentation of road scene image. The network uses SK attention mechanism to improve the feature fusion module, adjust the feature weights, and optimize the training process by image enhancement and hyperparameter adjustment. Through experiments on cityscape and other datasets, our method achieves the best segmentation results based on subjective visual effects and objective performance evaluation indicators in the comparison network, among which pixAcc reaches 91.2% and mIoU reaches 79.8% on cityscape dataset. It can be seen that the semantic segmentation effect of our method on road scene image is significantly improved. However, there are two specific limitations to note. Firstly, our method's performance may be compromised in extremely challenging conditions, such as heavy occlusions or severe weather effects, which were not extensively tested in this study. Secondly, while our model achieves high accuracy on urban road scenes, its applicability to rural or less structured environments remains uncertain and may require further adaptation. Future work should focus on developing adaptive models that can learn from real-time data, ensuring robustness in diverse scenarios. Overall, our proposed method not only improves accuracy but also sets the foundation for future innovations in autonomous driving and intelligent transportation systems.

REFERENCES

[1] S Minaee, Y Boykov, F Porikli, et al. “Image segmentation using deep learning: A survey”. IEEE transactions on pattern analysis and machine intelligence, vol.44,2021, pp.3523-3542.

[2] X Li, J Zhang, Y Yang, et al. “Sfnet: Faster and accurate semantic segmentation via semantic flow”. International Journal of Computer Vision, vol.132, 2024, pp.466-489.

[3] D Feng, et al. “Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges”. IEEE Transactions on Intelligent Transportation Systems, vol.22, 2020, pp.1341-1360.

[4] Y Wang, J Zhang, Y Chen, et al. “Automatic learning-based data optimization method for autonomous driving”. Digital Signal Processing, 2024, pp.104428.

[5] A de Silva, R Ranasinghe, A Sounthararajah, et al. “Beyond Conventional Monitoring: A Semantic Segmentation Approach to Quantifying Traffic-Induced Dust on Unsealed Roads”. Sensors, vol.24, 2024, pp.510.

[6] A Alzu'Bi, L Al-Smadi. “Monitoring deforestation in Jordan using deep semantic segmentation with satellite imagery”. Ecol. Informatics, vol.70, 2022, pp.101745.

[7] M Wieland, S Martinis, R Kiefl, et al. “Semantic segmentation of water bodies in very high-resolution satellite and aerial images”. Remote Sensing of Environment, vol.287, 2023, pp.113452.

[8] H Zhang, B Han, et al. “Slimmer: Accelerating 3D Semantic Segmentation for Mobile Augmented Reality”.2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), IEEE, 2020, pp.603-612

[9] S Afzal, IU Khan, I Mehmood, et al. “Leveraging Augmented Reality, Semantic-Segmentation, and VANETs for Enhanced Driver's Safety Assistance”. Computers, Materials & Continua, vol.78, 2024.

[10] D Zhang, L Zhang, J Tang. “Augmented FCN: rethinking context modeling for semantic segmentation”. Science China Information Sciences, vol.66, 2023, pp.142105.

[11] Li L, Dong Z, Yang T,et al. “Deep learning based automatic monitoring method for grain quantity change in warehouse using semantic segmentation”.IEEE Transactions on Instrumentation and Measurement, vlo.70, 2021, pp.1-10.

[12] F Abdullah, A Jalal. “Semantic segmentation based crowd tracking and anomaly detection via neuro-fuzzy classifier in smart surveillance system”. Arabian Journal for Science and Engineering, vol.48, 2023, pp.2173-2190.

[13] Y Wang, Y Shen, B Salahshour, et al. “Urban flood extent segmentation and evaluation from real-world surveillance camera images using deep convolutional neural network”. Environmental Modelling & Software, vol.173, 2024, pp.105939.

[14] J LONG, E SHELHAMER, T DARRELL. “Fully convolutional networks for semantic segmentation”. IEEE Computer Society, 2017, pp.3431-3440.

[15] V BADRINARAYANAN, A KENDALL, R CIPOLLA. “Segnet:a deep convolutional encoder-decoder architecture for image segmentation”. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.39, 2019, pp.2481-2495.

[16] C Kyunghyun, et al. “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation.” Conference on Empirical Methods in Natural Language Processing , 2014.

[17] H Noh, S Hong and B Han, "Learning Deconvolution Network for Semantic Segmentation," 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1520-1528.

[18] O Ronneberger, P Fischer, and T Brox. “U-net: Convolutional networks for biomedical image segmentation”. In International Conference on Medical image computing and computer-assisted intervention, 2015, pp.234–241.

[19] H Zhao, J Shi, X Qi, et al. “Pyramid scene parsing network”. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp.2881-2890.

[20] K He, et al., "Deep Residual Learning for Image Recognition". 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp.770-778.

[21] L C Chen, G Papandreou, I Kokkinos, et al. “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs”. arXiv, 2014.

[22] L C Chen, G Papandreou, I Kokkinos, et al. “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs”. IEEE transactions on pattern analysis and machine intelligence, 2017, pp.834-848.

[23] L C Chen, G Papandreou, F Schroff, H Adam. “Rethinking atrous convolution for semantic image segmentation.” 2017, arXiv:1706.05587.

[24] L C Chen, Y Zhu, G Papandreou, et al. “Encoder-decoder with atrous separable convolution for semantic image segmentation”. Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

[25] X Li, W Wang, X Hu and J Yang, "Selective Kernel Networks". 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 510-519.

[26] M Cordts, M Omran, S Ramos, et al. “The Cityscapes Dataset for Semantic Urban Scene Understanding”. IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp.3213-3223.

[27] Z Zhang, M Sabuncu. “Generalized cross entropy loss for training deep neural networks with noisy labels”. Advances in neural information processing systems, vol.31, 2018.